School of Public Policy, Central European University
Academic Year 2015–2016

Institut Barcelona d'Estudis Internacionals
Academic Year 2017–2018

Machine Learning for Public Policy Making

How to Use Data-Driven Predictive Modeling for the Social Good

Dissertation submitted by

FABIAN STEUER

in partial fulfillment of the requirements for the degree of

ERASMUS MUNDUS MASTER IN PUBLIC POLICY

Supervisors

Michael T. Dorsch

Jacint Jordana

July 2018

DECLARATION

I hereby certify that this dissertation contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I hereby grant to IBEI and the Mundus MAPP Consortium the non-exclusive license to archive and make accessible my dissertation in whole or in part in all forms of media, now or hereafter known. I retain all ownership rights to the copyright of the dissertation. I also retain the right to use in future works (such as articles or books) all or part of this dissertation.

Fabian Steuer

Hannover, 15 July 2018

ABSTRACT

Machine learning gives computers the ability to learn from data without being explicitly programmed. Due to its excellent prediction abilities, it has recently gained traction in economics, statistics and social sciences. Real-world problems machine learning has been applied to include predicting the probability that individuals commit crimes, targeting hygiene inspections by data-mining online restaurant reviews or estimating poverty levels based on satellite imagery. In this thesis I explore how machine learning can help to solve such and other prediction problems in public policy making and what challenges it faces. My goal is to bring the two fields closer together as most public policy makers likely do not even know that they face prediction problems that machine learning can help solving. After an introduction to prediction problems, I give an overview of how machine learning works and explain under what circumstances machine learning can be used for data-driven predictive modeling for the social good. A case study about predicting hygiene violations in restaurants illustrates the lessons learned and allows to get an idea of what applying machine learning looks like in practice. I then look into the challenges and limitations that machine predictions face in public policy making. Besides the fundamental limits of prediction, these range from technical and human challenges to ethical and legal issues due to biased predictions, black-box algorithms and questions of responsibility.

CONTENTS

FIGURES

TABLES

1 INTRODUCTION

Judges in the United States have to make a tough decision. When somebody is accused of having committed a crime, a judge has to decide where the suspect has to await trail: in freedom being granted bail, or in prison. By law, this decision must only depend on the prediction what the defendant would do if released (Kleinberg et al. 2017). Will she not show up to the trial? Or even commit another crime? Human judges have made these predictions for centuries, but unfortunately their predictions are not always correct. As more than 10 million people are arrested per year in the United States (U.S. Department of Justice 2016), these mispredictions sum up, with negative consequences for both defendants and society (Kleinberg et al. 2017). Recently though, a new player has emerged that could significantly improve how bail decisions are made. This new player is called machine learning. Machine learning gives computers the ability to learn from data, to create statistical models that capture patterns in data, and to make predictions based on the patterns. In the bail decision context, machine learning models trained on data about defendants arguably make more accurate predictions than human judges. Recent research found that through using such models "crime can be reduced by up to 24.8% with no changes in jailing rates, or jail populations can be reduced by 42% with no increases in crime rates"(Kleinberg et al. 2017). If this is true, the potential gains for defendants and society could be immense.

Bail decisions are just one example of prediction problems in public policy making. Big data, more powerful computers and advancing algorithms have decreased the cost of prediction and led to an increasing number of questions being reframed as prediction problems (Agrawal 2018). In the private sector, machine predictions are routinely used to obtain credit ratings, recommend books to online shoppers, or power self-driving cars. At first sight, none of these problems seems to be about prediction per se, but given the right approach, all of them can be phrased as prediction problems. It is not only the private sector that can profit from better prediction, but public policy makers also face many similar problems that can be solved with machine learning.

In one sentence, machine learning gives "computers the ability to learn without being explicitly programmed" (Munoz 2014). It is a subfield of artificial intelligence, located somewhere between computer science, statistics and adjacent disciplines (Jordan and Mitchell 2015). Machine learning and big data have recently gained traction in economics, statistics and quantitative social sciences (Taylor, Schroeder, and Meyer 2014; Einav and Levin 2014; Schroeder 2014). Machine learning and big data go hand in hand and enable many new insights in these disciplines. Better predictions are one of them. Given enough training data, machine learning models can often yield better predictions than humans or other models. How to use them for the social good, just like the challenges and limitations that this data-driven approach to predictive modeling faces, are what we will discuss in this thesis.

One of the foremost challenges that machine predictions face in public policy making is one of knowledge. The communication between machine learning experts, public policy makers and other domain experts is not always easy (Rudin and Wagstaff 2014). Most public policy makers probably do not even know that they face prediction problems that machine learning can help solving. This thesis aims at bringing the two fields closer together. There is no guide yet that gives an overview of how machine learning can solve prediction problems in public policy making and offers a concrete example in form of a case study. This thesis aims at filling this gap. There is vast literature on machine learning and public policy making respectively but not much in the intersection between the two fields. This thesis is intended for practitioners at any level of public policy making who face prediction problems on a regular basis. Public policy making in this sense is broadly defined because prediction problems exists in many different areas of the field. Anybody working for the social good instead of private interests is regarded as a public policy maker in this thesis. Acknowledging the often non-technical educational background of public policy makers, this thesis tries to find a balance between the technical language necessary for understanding machine learning and keeping technical details to a minimum. There is no dedicated literature review in this thesis, but instead the whole thesis draws from the available literature, somewhat similar to what a review article would do. For topics that we

cannot discuss in sufficient detail in this thesis, the literature references should give the interested reader a good starting point to dig deeper.

This introduction is followed by six more chapters. The following two explain what prediction problems are and give an overview of the machine learning workflow. They should provide the reader with sufficient knowledge about data-driven predictive modeling to follow the rest of this thesis. Chapter 4 contains a case study that illustrates what solving a prediction problem in public policy making looks like in practice. We will apply machine learning to predict which restaurants in a city likely violate hygiene regulations. Chapter 5 generalizes from the lessons learned before and explains for what kind of prediction problems in public policy machine predictions are suitable and what advantages machine learning offers over other methods. Chapter 6 then counterbalances the argument for machine learning. It goes into the challenges and limitations that machine predictions face in public policy making and when machine learning is not an appropriate method to solve prediction problems. The last chapter concludes with an overview the most important points.

2 PREDICTION PROBLEMS

Prediction problems are often seen as being mainly a concern of the private sector and not the public one. Private companies use predictive methods to recommend products, to predict machine failure or to recognize images. Even driving a car has been turned into a prediction problem by having computers predict how a human driver would react in a given situation (Agrawal 2018). At first sight, these are very different challenges, so what exactly are prediction problems? Prediction problems can be best understood when contrasting them with causal inference problems, which public policy makers should be familiar with. Most public policy makers likely ask two types of causal questions on a regular basis (Gelman 2010):

1. What is the effect of X? What is the effect of better teachers on test scores, the effect of speed limits on car accidents, the effect of subsidies on consumption? These questions look forward in the causal chain and ask for the effects of causes.
2. What causes Y? Why do people vote for populist parties? Why are some countries richer than others? Why is the banking system at the brink of collapse? These questions look backwards in the causal chain and ask for the causes of effects.

Both types of causal questions are commonly asked and well-known to public policy makers. They raise fundamental issues about the connection between policy interventions and impact, which every public policy maker should care about. Consequently, there is a lot of research being done in these areas and many methods to address these questions have been developed. Randomized controlled trials, synthetic control methods, regression discontinuity designs, instrumental variables and even traditional regression analyses are all used to answer these types of causal questions.

But causal questions are not the only ones the public policy makers have to answer. There are also prediction problems, for which causal inference is not necessary (Kleinberg et al. 2015). In the case of prediction problems, we do not ask causal "what is the effect of" or "why" question like the ones above but simply factual "what" or "how many" questions. What will the weather be like tomorrow? How many new

students will sign up for school next year? Where will most crimes be committed in a city? Answers to such questions are important for the wide range of public policy makers who have to deal with them on a regular basis. The local traffic authorities need to know about the weather, the ministry of education about the students, and the police about the crime prediction. Some of these questions, such as the question for the weather, have complex mathematical models as their answer. Such a model is informed by the causal factors which we know to influence the weather, but the model's purpose is not to derive the causal factors themselves. Other questions, like the one for the number of crimes committed in a specific area, might be based more on the experience of police officers and simple estimates. There is a common pattern in all prediction problems though. Put simply, prediction "is about using information you have to generate information you don't have" (Agrawal, Gans, and Goldfarb 2018). It is about using the known causes of an effect to predict the effect we are interested in. To add some details to this general definition, we can distinguish prediction problems along several dimensions.

The first dimension is temporal. As humans cannot see into the future, the information and data we have is always from the past. But predicting something often means predicting the future. In the bail prediction example this is clearly the case. A judge has to predict how likely a defendant will commit a future crime or not show up for the trial. Predictions do not necessarily have to be about the future though. The information we do not have can also be about the past or the present. Take surveys for example. Economic or social surveys are only conducted occasionally, not every year in every region. Consequently, there are data gaps because we do not know the results a survey would have delivered in a year it was not conducted. Filling these data gaps, for example through imputation or interpolation methods is an example of predicting the past. When it comes to predicting the present, many economic institutions use so-called nowcasting models to predict the current state of the economy (Mol et al. 2015; Taylor, Schroeder, and Meyer 2014). The problem that these nowcasting models solve is that many economic indicators can only be calculated long after the time period they refer to is over. It simply takes some time to gather and process all necessary data to determine the state of the economy at some point in time. Nowcasting models provide

speedy approximations to the indicators of interest and can provide policy makers with timelier information.

What kind of information can be predicted? This is another dimension along which we can distinguish prediction problems from each other. Lifting this question to a more abstract level leads to the distinction between regression and classification problems. In the examples above the goal is to predict numbers such as the centimeters of snowfall, the number of students or the crime rates in different areas. These examples are instances of regression problems, in which the goal is to predict a continuous numeric variable. Since the numbers to be predicted in our examples can have any numeric value (in a range of sensible values) they are regression problems. Other examples of regression problems involve predicting probabilities on a scale from 0 to 1 or bringing items in a certain order (by predicting a continuous numeric value and then ordering the items according to this value). In classification problems, on the other hand, the goal is to predict a class. A simple two-class classification problem, for example, is fraud detection. The European Commission uses fraud detection software for their public tenders to ensure that bidders do not copy from other bidders or other illicit sources.[1] If this software classifies submitted documents into fraud and non-fraud documents, the algorithm has a two-class classification problem to solve. Three or more classes are possible too, for example non-fraud, maybe-fraud and fraud. This three-class problem is ordered because the three classes can be put into some natural order, ranging from a small likelihood of fraud to very likely fraud. Other problems require unordered classification, such as segmenting the types of patients in a public hospital into disease categories to match them to the hospital's departments. This classification problem has no natural ordering and is thus unordered.

The boundaries between these classes of prediction problems are not always clear-cut. The number of cars on the road of a city on a given day, for example, strictly speaking is not a continuous variable because the number of cars is always a whole number. Barring accidents, we cannot have 42 and a half cars on the roads. But given that there

[1] Based on my own experience while working for the European Commission's Secretariat-General.

are so many possible different numbers of cars, this is more a regression than a classification problem. Where the past ends and the present or the future starts on the arrow of time is not always clear either. Nonetheless, it is useful to keep these dimensions in mind when reasoning about prediction problems. Considering whether there is any unknown information in the past, present or future, be it continuous or more coarse-grained in nature, which could be useful for a public policy maker is a good point to get started with predictive modeling for the social good.

Coming back to the distinction at the beginning of this chapter, causal inference and prediction go hand in hand. Prediction is most useful when we already know about relevant causes and effects. If the weather forecast predicts snow, the traffic authorities prepare to send out their snowplows; if the number of students will likely increase, the ministry of education hires more teachers; if crimes are likely to happen in a certain area, the police might send more officers to patrol the area. In all these cases some measure (snowplows, hiring teachers, deploying police officers) is known to have some desired effect in some environments, but not in others. Snowplows are only useful when it snows, hiring more teachers makes most sense when we expect more students, deploying police can deter crime but is a waste of public funds if there are no crimes or other relevant problems. In these cases, policy makers need to predict what the environment will be at a relevant point in time, so that they can react to it appropriately using the information they have about the effects and causes of certain measures. Public policy makers make such predictions all the time, often based on their intuition and experience. Machine learning is a way to make these predictions more data-driven and accurate, as we will see in the following chapters. This being said, machine learning can even help to improve the causal inference techniques mentioned above. Randomized controlled trials, synthetic control methods, regression discontinuity designs and instrumental variables all involve an element of prediction. As these methods are only known to some experts in public policy making, how exactly machine learning can help improving them is explained in the appendix on page 48.

3 MACHINE LEARNING

How do public policy makers solve prediction problems? One approach is to rely on human experience and intuition. It might work quite well, for example, to deploy police officers based on where the police department thinks crimes are most likely. After all, this is what many departments have always been doing. But it might also work not so well. Would it not be great if we could use existing data on crime rates and other existing information to improve the police department's crime predictions? This is where machine learning comes into play. Machine learning uses information we have to tell us something about information we do not have. It is about using data to make predictions. The more data, the better a machine learning algorithm can learn from the data. That is why the method is called machine *learning*. Just like humans learn from experience, machine learning gives computers the ability to learn from data. This chapter provides a high-level overview of what machine learning is and what steps training a machine learning model involves.

Machine learning is best understood when contrasted with traditional human data modeling. Many public policy makers should be familiar with linear regression models, simple statistical models that describe linear relationships between independent and dependent variables. Linear regressions and other traditional data models are both similar to and different from machine learning models. That is why it is worth having a closer look at the similarities and differences between these two cultures of data modeling (Breiman 2001). Both human data modeling and machine learning try to model relationships in data. Both use statistical models to describe the relationship between independent and dependent variables in the data. The crucial difference between human data modeling and machine learning is how the model that describes the relationship is chosen. In human data modeling the human modeler chooses a stochastic model such as a linear regression model, which is then fit to the data. This reduces the problem of fitting an almost arbitrary function to the problem of finding a limited number of parameters that maximize the fit of a given functional form to the data (Breiman 2001). In the case of linear regressions, these parameters are

the axis intercept and the slope. Finding the parameter values that maximize the fit of the model to the data is relatively easy using such methods as least squares.

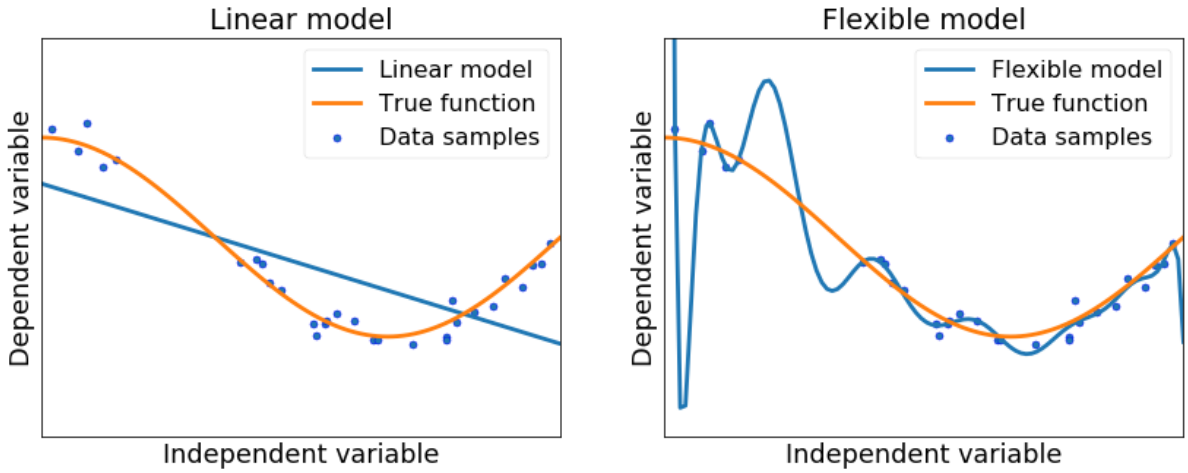

Figure 1: A linear model and a more flexible model fit to the same dataset. The linear model underfits the data as it is not flexible enough to fit the curved relationship between the independent and the dependent variable. However, the more flexible model is too flexible and overfits the data. Following the sampled data points too closely leads to a very wiggly curve that does not fit the true data-generating function very well. Consequently, the ideal model would be more flexible than the linear model on the left but less flexible than the too flexible model on the right.

In machine learning a human also chooses a particular algorithm to model the relationship between independent and dependent variables. In contrast to human data modeling, however, machine learning algorithms are able to learn very flexible functional forms from data without humans having to define them explicitly. This gives machine learning a crucial advantage: machine learning algorithms are able to fit complex models on their own. Choosing a machine learning algorithm restricts the possible relationships that can be learned way less than a human-chosen model such as a linear model would do. The price to be paid for the simplicity of a linear model (or another data model that humans are able to define using simple formulas) is that the model will not fit the data very well if the relationship between independent and dependent variables is not linear (or generated according to whatever relationship the human modeler has chosen). Of course, it is possible to add nonlinear transformations of the independent variables to a linear regression equation. But this still restricts the relationships that can be learned to the limited set of functional forms chosen by the human modeler. Machine learning, on the other hand, enables the computer to

automatically fit more flexible models to describe the relationship between independent and dependent variables. This makes model fitting somewhat harder because there are far more parameters in machine learning models than there are in such simple models as linear regressions. But if machine learning algorithms are given enough training data, they can fit more complex relationships between independent and dependent variables than a data model defined by a human could ever do.

Admittedly, the distinction between machine learning and human data modeling is blurred. Even fitting a simple linear model using least squares strictly speaking is machine learning. Although a human chooses the linear model, it is the computer that fits the model to the data using least squares. Machine learning offers many more algorithms though, which can model more complex relationships in data. Unfortunately, using them comes at a cost. Just as a linear model can be too restricted to fit a complex relationship, machine learning models can be too flexible, so that they fit some data too well. This problem is illustrated in Figure 1. If a model is not flexible enough to fit the true relationship between independent and dependent variables, one speaks of underfitting. The opposite, when a model is that flexible that if follows every data point in the sampled dataset that the model was fit to, is called overfitting. The art in machine learning is to find the balance between the two. A model should neither be that restricted that it underfits the true relationship between independent and dependent variables, nor that flexible that it overfits a given data sample.

How to use machine learning for prediction in practice? Figure 2 shows the high-level workflow of creating and using a machine learning model[2], consisting of data preparation, model training and making predictions. Let us have a look at what each of these steps entails. The following pages should give public policy makers a basic understanding of machine learning. More details on the machine learning workflow can be found in the literature. James et al. (2007) provide a good introduction, while

[2] Technically speaking, the figure illustrates the workflow of training a model through supervised machine learning, which is different from other types of machine learning such as unsupervised learning. To keep technical terminology to a minimum, "machine learning" in this thesis generally refers to "supervised machine learning" as this is the most common type of machine learning and the one that is most useful for making predictions (Jordan and Mitchell 2015). Explanations of how other types of machine learning work can be found in the literature.

Hastie, Tibshirani and Friedman (2009) offer a detailed overview of different machine learning methods.

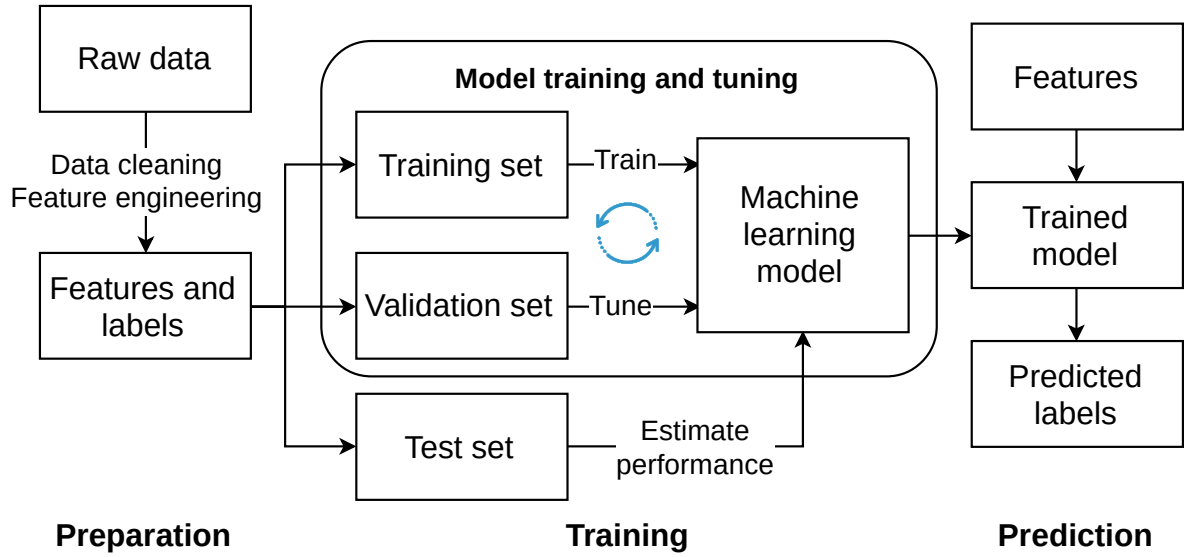

Figure 2: Overview of the machine learning workflow from data preparation and model training to prediction. Further details of this workflow can be found in the machine learning literature.

3.1 Data preparation

The machine learning workflow starts with raw data that we believe to be useful for predicting the outcome we are interested in. Raw data rarely comes in a form that is directly useful for machine learning. Usually, data has to be cleaned, data quality problems have to be solved, and different data sources have to be consolidated into a single dataset. After this step, a dataset usually looks similar to a table with variables in columns and observations in rows. In machine learning the variables are also called features, which explains why the next step in the workflow is referred to as feature engineering. Feature engineering means to extract higher-level features from raw variables (that is, lower-level features) in the data to give a machine learning algorithm more useful information. Coming back to the bail prediction example, imagine we want to predict whether a defendant is likely to appear before court if released on bail. We gather data on past bail decisions and what the defendants did if they were released. Once we have cleaned and consolidated the data into a single dataset, one feature in our dataset is the charge on which the defendant was arrested. As Table 1 illustrates, a higher-level feature based on the arrest charge would be an additional

feature that describes whether a defendant is charged with a violent crime or not. If we believe the violent crime category to be useful to predict if a defendant would appear before court if released on bail, this feature engineering step could add valuable information to our dataset. In addition to the features, our dataset also needs to contain information on the outcome we want to predict, so that a machine learning algorithm can learn the patters that lead to a defendant's failing to appear before court. In machine learning this target variable that we are interested in is called a label. At the end of the data preparation process, there should be a set of relevant features and a single label for every observation in the dataset.

	Features							**Label**
ID	**Age**	**Gender**	**Race**	**Arrest county**	**Arrest charge**	**Violent crime**	**Prior arrests**	**FTA**
1	28	Male	White	Bronx	Murder	Yes	Drugs	Yes
2	35	Female	Hispanic	Queens	Robbery	Yes	-	No
3	21	Male	Black	Brooklyn	Fraud	No	Guns	No
…	…	…	…	…	…	…	…	…

Table 1: Made-up examples of features that could be used for predicting whether a defendant would appear before court if released on bail. FTA is the label to be predicted and stands for "failure to appear". Some of the features such race might actually not be included in a predictive model because it has become politically unacceptable to use them (Angwin et al. 2016).

3.2 MODEL TRAINING

Once the data is clean and there are meaningful features and a label for every observation in the dataset, we can start training the actual machine learning model. To begin with, the data is split into three smaller, non-overlapping datasets: a training set, a validation set and a test set. As the names suggest, a machine learning model is trained on the training set, validated using the validation set, and evaluated on the test set. This firewall principle ensures that none of the data used to train the model is used to evaluate it (Mullainathan and Spiess 2017). First, a machine learning algorithm takes the features and the labels in the training set and fits a statistical model to the data. Just like a linear regression finds the optimal linear relationship between the independent variables and the dependent variable, machine learning tries to find the

relationship between the features and the label – which might be much more complex than a simple linear relationship. For example, maybe defendants charged with violent crime are less likely to appear before court. Or young defendants, arrested in Brooklyn with a prior arrest on gun charges. There is no theoretical limit on how complex the relationship between the features and the label can be. Given enough data, machine learning algorithms are able to find these patterns automatically, without a human having to specify a limited number of functional forms to be fit to the data.

Machine learning faces a challenge though. Given a sufficiently flexible model, it is possible to fit any dataset arbitrarily well. The graph on the right of Figure 1 shows that the more flexible a model is, the better it can fit the samples in a dataset. Unfortunately, this is not exactly what we want. The goal in machine learning is to make predictions using previously unseen data, not to fit the given data as closely as possible. A machine learning model needs to generalize beyond the dataset it was trained on (Domingos 2012; Mackenzie 2015). This is where the validation set comes into play. Machine learning models have many parameters that are fit to a dataset, just like the intercept and the slope in a linear regression. But in addition to these normal model parameters, machine learning models also have hyperparameters. Put simply, the hyperparameters of a model control how flexible the model is in fitting the data. After initially training a machine learning model on the training set, we can use the validation set to validate how accurate the model is in predicting previously unseen labels. We let the machine learning model make predictions using the features in the validation set and compare the predicted labels to the true labels in the validation set. This gives us an idea of how well the model generalizes. If the model fails to generalize beyond the training set (if the predictions on the training set are significantly better than the predictions on the validation set), it is likely overfitting. To avoid overfitting, we can adjust the hyperparameters of the model to make it less flexible and generalize better (Athey and Imbens 2016). Models that do not generalize well are unlikely to make accurate predictions when they are deployed in a real-world context, so this is a crucial step in training a machine learning model. Once we are satisfied with how well the model generalizes from training to validation set, we stop training the model. The test set is then used to give us a final unbiased estimate of the model's predictive

performance. Comparing the labels that the model predicts for the test set data to the true labels in the test set allows us to evaluate how well the model is able to make predictions for data that has not been used during training.

3.3 PREDICTION

At this point we hopefully have a promising machine learning model. We can now start making predictions in a real-world context. Taking the bail prediction example, we now want to use the trained machine learning model to predict how likely a new defendant would be to appear before court again if released on bail. Making predictions using a trained machine learning model is easier than training the model in the first place. We simply use the characteristics of a defendant as features, based on which the model predicts the probability that the defendant will appear before court. The prediction might then be used to make the bail decision, or at least to give the judge some additional information to make her decision.

4 CASE STUDY - PREDICTING HYGIENE VIOLATIONS

Virtually every city in the world needs to ensure food safety. The food offered in restaurants and other food places should be safe for consumption. Many cities employ hygiene inspectors for this purpose, who check restaurants to identify and solve hygiene issues. But there is a problem: cities usually do not have enough inspectors to visit every restaurant sufficiently often, so some hygiene violations go unnoticed and threaten public health. This is the setting for this case study. Our goal is to better allocate hygiene inspectors to restaurant to catch as many hygiene violations as possible. Ideally, hygiene inspectors focus on the restaurants that most likely commit hygiene violations – which we can predict using machine learning. Coming back to the difference between causal inference and prediction, the predicted risk that a hygiene violation occurs is useful information for food authorities because they know that hygiene inspectors are a useful causal means to combat hygiene violations.

In this case study I use a machine learning model trained on past data from the City of Seattle to predict future hygiene violations. The data used comes from the paper "Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews" by Kang, Kuznetsova, Luca and Choi (Kang et al. 2013).[3] The dataset combines hygiene inspection scores from the City of Seattle with restaurant reviews posted on Yelp, a website for crowd-sourced reviews of local businesses.[4] Kang et al. (2013) scraped reviews written between 2006 and 2013 for restaurants in Seattle from Yelp and combined them with the hygiene inspection records of the city. This resulted in a dataset containing 13,299 inspections of 1,756 restaurants with 152,153 Yelp reviews. Analyzing this dataset by hand would take very long, but for machine learning standards it is a rather small dataset.

Before training the machine learning model, we need to understand the context better. How exactly are hygiene inspection scores in Seattle generated? When an inspector inspects a restaurant in Seattle, she assigns an inspection penalty score to quantify how

[3] The data is available for download at http://www3.cs.stonybrook.edu/~junkang/hygiene/
[4] https://www.yelp.com/

well the restaurant complies with public health and food regulations. Figure 3 shows the distribution of the scores in the Kang et al. (2013) dataset. The higher the score, the worse. However, low positive inspection scores do not necessarily mean that there are dangerous hygiene problems. Minor violations such as improper labelling are noted by the inspectors and lead to higher inspection penalty score but are not too big a danger for public health (Kang et al. 2013). To simplify the prediction problem, I therefore do not try to predict the inspection score, but whether a score is high enough to be classified as severe hygiene violation. This reduces the prediction problem from a regression problem to a two-class classification problem. Since there is no threshold above which hygiene violations are officially severe, I define approximately the worst 10% of all hygiene violations as severe violations. This results in a penalty score threshold of 33, from which on violations are classified as severe. Exactly 10.3% of all hygiene scores found in the dataset fall into this category.

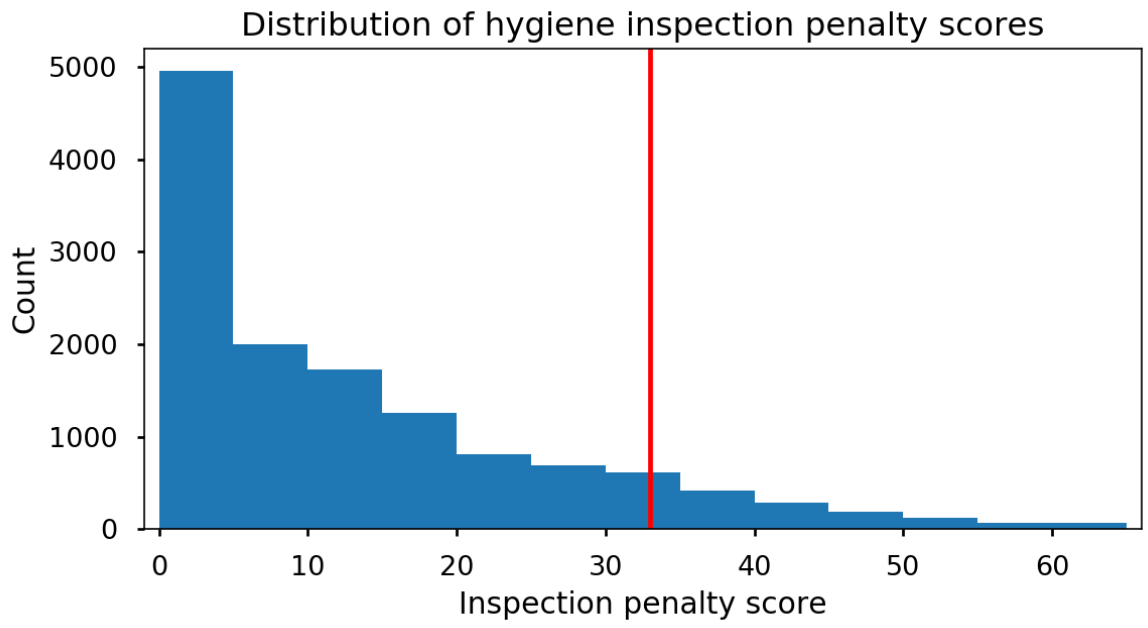

Figure 3: Distribution of hygiene inspections penalty scores resulting from 13,299 hygiene inspections of 1,756 restaurants in Seattle between 2006 and 2013. The higher the penalty score, the worse. The red line marks the penalty score threshold of 33 used in the case study. Hygiene violations with a score equal to or greater than 33 are classified as severe. 10.3% of all violations in the dataset fall into this category.

Given that there are far more inspection scores than restaurants in the dataset, many restaurants have several inspection scores assigned to them at different points in time.

That is why I follow Kang et al. (2013) and define the *inspection period* of an inspection score as the period from the day after the previous inspection to the day of the inspection in question. For restaurants without any previous inspection, the first inspection period spans the last 6 months before the first inspection. Figure 4 shows the distribution of the lengths of the inspection periods according to this definition.

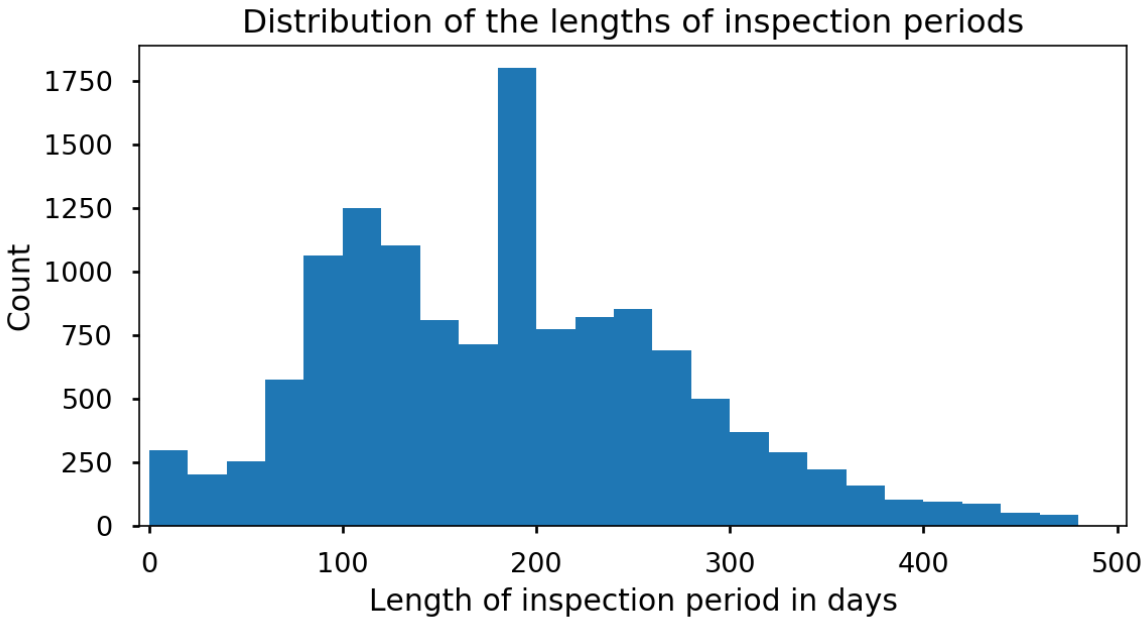

Figure 4: Distribution of the lengths of the inspection periods in the hygiene prediction dataset in days. The inspection period of an inspection score is defined as the period from the day after the previous inspection to the day of the inspection in question. For restaurants without any previous inspection, the first inspection period spans the last 6 months before the first inspection (hence the unusually high peak in the histogram around the period of 180 days).

The prediction problem we are facing now is to predict the outcome of the hygiene inspection at the end of each inspection period. This is a simple two-class prediction problem. Are the violations found in a hygiene inspection severe or not? What data can help us to answer this question? Table 2 shows the features used to train the machine learning model in this case study. The easiest way to imagine them is a huge table where each inspection period is a row and each feature is a column.

Data	Explanation
ZIP Code	The ZIP Code of the restaurant.
Cuisines	The cuisines offered in the restaurant according to Yelp, such as Japanese, Mexican, Pizza, Sandwiches etc.
Length of the inspection period in days	The inspection period of a hygiene inspection ranges from the day after the previous inspection until the day of the inspection in question. The first inspection period of a restaurant spans the six months before the first hygiene inspection of that restaurant.
Number of reviews	The number of reviews of a restaurant that users posted on Yelp during the inspection period.
Average review rating	The average rating of the reviews of a restaurant posted during the inspection period (ranging from one to five stars).
Number of negative reviews	The number of reviews of a restaurant with a rating below or equal to three stars that users posted on Yelp during the inspection period.
Average previous inspection penalty score	The average of the hygiene inspection penalty scores assigned to a restaurant before the inspection in question (zero if there has been no previous inspection).
Previous inspection penalty score	The hygiene inspection penalty score assigned to a restaurant in the last inspection before the inspection in question (zero if there has been no previous inspection).
Review text	The concatenated texts of all reviews posted during the inspection period.
Inspection penalty score	The inspection penalty score assigned to a restaurant in the inspection period in question. The goal is to predict if this score is equal to or greater than 33, in which case a hygiene violation is labeled as severe.

Table 2: Features and label contained in the dataset used in the hygiene violation case study.

A significant part of the data preparation necessary for this case study has been done by Kang et al. (2013), who scraped restaurant information and review texts from Yelp and combined them with hygiene inspection scores from the City of Seattle. But there is some feature engineering left to do. Training machine learning algorithms requires numeric data. Most of the features in Table 2 such as the length of the inspection period in days, the number of reviews, or the ZIP code are numeric, so they have the correct format. The cuisines feature and the hygiene violation label can easily be converted to numeric features by coding them as dummy variables. But how to convert the review texts into numeric data?

Figure 5: A Yelp review that might indicate hygiene problems in a restaurant. Some personal information has been removed from the review for privacy reasons.

As Figure 5 shows, a Yelp user is free to write virtually anything in a review, ignoring grammar and spelling rules or even inventing new words. Machine learning is able to find pattern even in this messy data, but cleaning the text before using it as input into a machine learning algorithm likely yields better results. That is why I prepared the review texts in three steps, using standard algorithms[5]:

1) I removed all punctuation marks from the reviews and converted all words to lower case.

[5] The algorithms used can be found in Python's spaCy package at https://spacy.io/

2) I removed all stopwords from the reviews. Stopwords are words such as "the", "but", and "for" that are so frequent in the English language that they do not contain much information that is relevant for prediction.
3) I lemmatized all remaining words by removing all inflectional endings and converting them to their dictionary form. For example, "goes", "went" and "going" were all converted to "go".

After this preprocessing, numeric features can be extracted from the review texts. I used a statistic called term frequency-inverse document frequency (TFIDF) that I applied to the concatenated texts of the reviews in an inspection period. The basic idea behind TFIDF is to calculate how frequent a word is in the review text of a particular inspection period and divide this frequency by a number summarizing in what share of all review texts the word is used.[6] If we just calculated the frequency of a word in a review text without normalizing it using the frequency of that word in all review texts, words that are more common in the English language would be automatically overrepresented compared to less common words. TFIDF avoids this problem and should give a good idea of how frequent a word is in a review text compared to its frequency in all review texts. To provide the machine learning algorithm with more useful information, I did not only use single words (so-called unigrams) to calculate the TFIDF, but also applied it to word pairs and word triples (so-called bigrams and trigrams). Table 3 shows how these n-grams are extracted from a preprocessed example sentence.

Original sentence	**We enjoyed our food, although parts of it were burned.**
Preprocessed sentence	enjoy food part burn
Unigrams	enjoy, food, part, burn
Bigrams	enjoy food, food part, part burn
Trigrams	enjoy food part, food part burn

Table 3: The text preprocessing steps used in the hygiene violation case study applied to an example sentence. The preprocessing steps are removing all punctuation marks, converting all words to lower case, removing stopwords, and lemmatizing the remaining words.

[6] The exact explanation of how the algorithm works can be found at http://scikit-learn.org/stable/modules/feature_extraction.html#tfidf-term-weighting

I only added the TFIDFs of the 5,000 most frequent uni-, bi-, and trigrams to the existing dataset. The resulting tabular dataset still has 13,299 rows (one for each inspection period), but now 5,135 columns: one for each feature, including those in Table 2 (with the ZIP code and the cuisines encoded as dummy variables) plus the TFIDFs of the 5,000 most frequent uni-, bi-, and trigrams. The result is large dataset with 5,135 features for every inspection period. Traditional modeling techniques could have difficulties with such a dataset where the number of variables is in the same order of magnitude as the number of observations, but machine learning can handle it.

Having prepared the dataset, I used 80% of the dataset consisting of 10,639 inspection periods for the training and validation sets and the remaining 2,660 inspection periods for the test set. I then trained several machine learning models using the training and validation set. The model that achieved the best results was a so-called random forest classifier. How this and other machine learning algorithms work is beyond the scope of this thesis, but many good explanations can be found in the machine learning literature. In any case, often the choice of a particular machine learning algorithm is less important than good feature engineering and having relevant data to begin with. In general, good data beats cleverer algorithms (Domingos 2012).

How well does the trained random forest model make predictions? For evaluating its predictive performance, I used the so-called area under the receiver operating characteristic curve (AUC), a metric that is common for evaluating predictions in machine learning. Our random forest model reaches on AUC of 0.7, which is shown in Figure 6. Despite its complicated name, AUC has an intuitive explanation. If you randomly pick an inspection period from all inspection periods that end with a severe hygiene violation and another one randomly from all inspection periods that do not end with a severe hygiene violation, the AUC is the percentage of such randomly drawn pairs for which the model predicts a higher hygiene violation risk for the inspection period that ended in a severe hygiene violation. In other words, given a random pair of inspection periods, one of them with hygiene violation and the other one without, there is a 70% probability that our model ranks the inspection period with the severe violation higher in risk than the one without severe violation. This is a

success for our model. A perfect model would rank every single one of such pairs correctly and have an AUC of 1, but a 70% probability of ranking correctly is significantly better than random guessing, which would result in a probability of only 50%. Our machine learning model seems to have learned something useful.

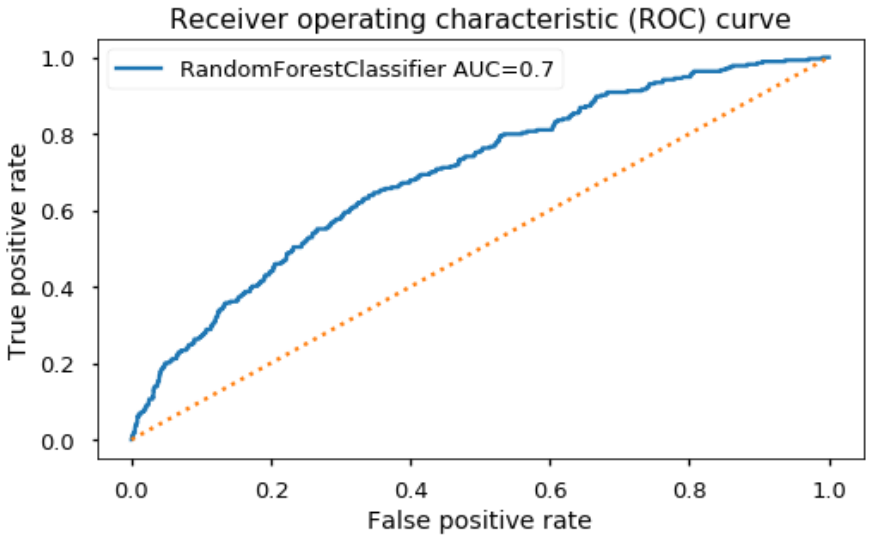

Figure 6: Receiver operating characteristic curve (ROC) for the random forest classifier used in the hygiene prediction case study. The statistical literature provides exact definitions of the true and false positive rates shown in this plot. For the purpose of this case study, it is merely important that the classifier reaches an area under the receiver operating characteristic curve of AUC = 0.7. This means that given a random pair of inspection periods, one of them with hygiene violation and the other one without, there is a 70% probability that the model ranks the inspection period with the severe violation higher in risk than the one without. The dotted line indicates the worst ROC possible with an AUC of 0.5. The best ROC possible would have a value of 1 everywhere, which results in AUC of 1.

How useful is our model in practice though? Although the AUC is a common metric in machine learning, it does not tell us much about the real-world predictive performance of our model. To get a better idea of how well our model makes predictions, it makes sense to have a look at the different types of prediction errors the model makes. Table 4 shows the so-called confusion matrix, which compares the model's predictions with the actual labels in the test set. There are two possible kinds of prediction errors: false-positive and false-negative ones. False positive errors occur when a model predicts a severe hygiene violation, but there was no severe violation in reality. False negative errors are those where there was a severe hygiene violation in an inspection period, but a model predicted that there was none. As the confusion matrix shows, our model reaches an overall accuracy of 75.5% on the 2,660 inspection

periods in the test set. This means that 126 + 1,882 = 2,008 of the model's predictions were correct. This is not too bad, but still leaves 504 false positive and 148 false negative predictions, which add up to 24.5% of all predictions in the test set. Relying on this model alone for allocating hygiene inspectors to restaurants in Seattle is thus likely a bad idea. The predictions could be used to give hygiene inspectors additional information as to which restaurants are likely to commit hygiene violations, but as every fourth prediction is wrong it is probably not reliable enough on its own.

Accuracy: 75.5% (n= 2,660)		**Predicted label**	
		Severe violation	**Not severe**
Actual label	**Severe violation**	**126** (true positives)	**148** (false negatives)
	Not severe	**504** (false positives)	**1,882** (true negatives)

Table 4: Confusion matrix for the random forest classifier trained using the training and validation sets of in the hygiene violation case study. The predictions are evaluated on the 2,660 inspection periods in the test set, which comprises 20% of the full dataset.

Nonetheless, this case study shows that using data-based predictive modeling to improve the allocation of hygiene inspectors to restaurants can be a promising approach. Hygiene inspections are a suitable prediction problem for machine learning. There are repeated inspection decisions to be made on a regular basis; there are new data sources in the form of Yelp reviews that contain predictive patterns; and the hygiene violation predictions are easy to evaluate by sending an inspector to a restaurant. Since Yelp offers a common online platform not only for Seattle but also for many other cities, it would be easy to scale up the machine learning model created in this case study to use it in other regions as well. This would probably improve the predictive performance of the model and allow it to have real-world impact. Recent research, for example, found, that the City of Boston could be 30-50% more effective in allocating hygiene inspectors to restaurants if it used the winning algorithm from an online machine learning tournament (Glaeser et al. 2016). That is a huge win for budget-stripped city administrations and shows how machine learning can improve public policy making.

5 MACHINE PREDICTIONS

Solving prediction problems with machine learning results in machine predictions. We have seen that prediction problems exist in public policy making and how machine learning can be a viable approach to solve them. In this chapter we will have a closer look at when exactly machine predictions can be useful for public policy making and what advantages machine learning offers in such cases. We will also see some more examples of prediction problems in public policy making to which machine learning has already been applied.

5.1 SUITABLE PREDICTION PROBLEMS

Not all prediction problems in public policy making are suitable for machine learning. There is a number of conditions to be met for machine predictions to be useful. One of them is that the outcome to be predicted must be frequent enough to be captured by statistical methods. Public policy makers would like to know the outcome of many uncertain events, but some of them are so unique that data-based modeling is the wrong approach. Who will win the next war? Will a scandal thwart the plans of the governing party? What will be the next decision of a capricious head of state? Although these are important questions, machine learning cannot answer them because these events are rare, and each one is special in its own way. For machine learning, the outcomes to be predicted need to be sufficiently similar to leave behind similar patterns in data, which then can be picked up by an algorithm. This implies that the objective to be predicted must not be too complex. The hygiene violations in the case study can be detected by sending an inspector to a restaurant, which all have to comply with the same regulations. Bail predictions are also relatively straightforward because they mainly depend on the predicted probability that a defendant would appear before court if released on bail. But predicting other kinds of court decision is a much more difficult problem. Factors that can be predicted such as recidivism risk play a role in sentencing, but many other factors such as deterrence, retribution and remorse, which are very difficult to measure, are usually also taken into account. Consequently, sentencing is too complex a problem for machine learning

(Kleinberg et al. 2017). A related condition is that the predictions made by a model can be evaluated. Great predictive performance on the data collected to train an algorithm does not help much if its predictions in the real world are too inaccurate. We need to test if an algorithm's predictions are correct and generalize well. If an algorithm, for example, predicted that a defendant would not commit a crime when released, but the defendant does commit one, the prediction was obviously wrong. Unfortunately, it is not always that simple to evaluate predictions, as we will discuss in Chapter 6.

5.2 ADVANTAGES OF MACHINE LEARNING

What advantages does machine learning have over other data modeling approaches? Unsurprisingly, the main advantage of machine learning is that it excels at prediction. This ability stems from different factors. One factor is that machine learning can deal with high-dimensional data, with the number of variables being in the same order of magnitude as the number of observations. Machine learning is very good at recognizing informative patterns in data and connecting them to the prediction objective. A related advantage is that the data used in machine learning can come in a wide variety of formats. A simple one is just a big table in which the columns are variables and the rows observations. But as big data is often a byproduct of the many computer mediated transactions taking place in our modern world (Varian 2010), it can also include text, images, videos, sounds, sensor data and many more – in general, any information that is digital or can be digitized. Every transaction on the internet, for example, leaves digital traces behind, which can be very useful for public policy making. Some people also define big data in terms of the so-called three Vs: larger data volume, velocity and variety (Hassani and Silva 2015; Booz Allen Hamilton 2015). Besides being big, big data is usually available with a shorter time lag and in a more granular form than traditional data (Einav and Levin 2014; Jordan and Mitchell 2015). Data aggregated at country or regional level, which is still used a lot in public policy making, for example, underutilizes machine learning's capabilities as a lot of information is lost during the aggregation. Public policy making could get a lot of prediction ability alone by using new sources of granular big data instead of aggregated datasets. Machine learning's last advantage becomes apparent once a

model has been fit. In this case it is easy to scale the model up (Kleinberg, Ludwig, and Mullainathan 2016). Once trained, a model can make predictions on new data at close to zero marginal cost. All you need is a system that stores data and uses the model to make predictions. One should of course evaluate if the predictions based on the new data are accurate, but given that this is the case, scaling up a machine learning model is straightforward and promises less variable predictions than humans.

5.3 EXAMPLES FROM PUBLIC POLICY MAKING

What prediction problems are there in public policy making? Table 5 gives an overview of prediction problems in a variety of fields in public policy making and how they can be solved using data-driven models. It outlines the prediction objective, the data and methods used, and the results of the paper. Some papers use advanced machine learning techniques, whereas others use more traditional regression models. But in both cases machine learning is potentially a viable approach to create predictive models. As this table shows, machine learning and data-driven modeling can be used in many different settings in public policy making. The examples range from agriculture, education and public engineering to economic, tax and health policy and come from both economically developing and developed countries. Prediction problems play a role from the municipal to the international level and many different data sources, including satellite imagery, electronic health records and classroom data can be used to tackle them. The table illustrates how many different prediction problems there are in public policy making and how machine learning can help to solve them.

Seeing this table, one question remains: what ways are there for public policy makers to solve prediction problems with machine learning in practice? Even most quantitative public policy makers who are familiar with data modeling, probably have never applied a machine learning algorithm themselves because the method has only recently gained a lot of attention. Although quantitatively oriented policy makers could probably learn how to apply machine learning from books and (online) courses, a more likely solution is to cooperate with consultants who are experts in machine learning. Discussing the prediction problems at hand with them, while being aware of

the challenges and limitations that we discuss in Chapter 6, is probably the easiest way to solve a prediction problem through machine learning. If this is too costly, a cheaper approach would be to cooperate with volunteers in the data science community. Data scientist and machine learning experts are often quite open to work for free for projects that promote the social good. Machine learning competition platforms such as Kaggle[7] or DataKind[8] can be a good starting point for solving machine learning problems. These platforms publish datasets together with a specific prediction problem and invite the machine learning community to solve it. The best solution usually wins a prize. Browsing the completed prediction competitions on these websites gives a good idea of what kinds of prediction problems have already been solved using this approach.

	Paper	Method, data and goal	Result
Economic policy	Combining satellite imagery and machine learning to predict poverty (Jean et al. 2016)	Convolutional neural networks are trained on publicly available high-resolution satellite imagery to estimate local consumption expenditure and asset wealth in five African countries.	The inexpensive and scalable model can explain up to 75% of the variation in local-level economic outcomes. It could transform how poverty is targeted and tracked in developing countries.
	Predicting poverty and wealth from mobile phone metadata (Blumenstock, Cadamuro, and On 2015)	Automatic feature engineering and elastic net regularization are used on an individual's past history of mobile phone use to infer her socioeconomic status.	In regions where censuses and household surveys are rare, the method allows to gather inexpensive, localized and timely information (at a finer level than satellite imagery).
Agricultural policy	Random Forests for Global and Regional Crop Yield Prediction (Jeong et al. 2016)	Random Forests are compared to multiple linear regressions for their ability to predict crop yields of wheat, maize, and potato using climate and biophysical variables at global and regional scales.	Random Forests outperformed multiple linear regression benchmarks in all performance statistics, making them an effective and versatile machine learning method for crop yield prediction.

[7] https://www.kaggle.com/
[8] http://www.datakind.org/

Health policy	Predictive Modeling for Public Health: Preventing Childhood Lead Poisoning (Potash et al. 2015)	Logistic regressions, support vector machines and random forests are used to predict the risk of children being poisoned by lead in their homes in Chicago. Data comes from blood tests, home lead inspections, property value assessments and censuses.	The models allow the Department of Public Health to prioritize which households to target when trying to prevent lead poisoning before it occurs. This is a better method than waiting for blood tests to indicate poisonings after the fact.
	Prediction modeling using EHR data: Challenges, strategies, and a comparison of machine learning approaches (Wu, Roy, and Stewart 2010)	Logistic regressions, support vector machines and boosting are used on data from electronic health records (EHRs) to detect heart failure before the actual date of clinical diagnosis.	The models are able to predict heart failure more than 6 months before the actual clinical diagnosis reasonably well. This means that a patient's health history can be used to predict future illnesses and target interventions.
Engineering	Water pipe condition assessment: a hierarchical beta process approach for sparse incident data (Li et al. 2014)	Bayesian nonparametric learning and existing infrastructure data are used to predict the failure probability of water pipes in a city to establish a ranking for inspections.	Experimental results show that the model does better than current best practice methods, leading to substantial savings on reactive repairs and maintenance.
Public hiring	Productivity and Selection of Human Capital with Machine Learning (Chalfin et al. 2016)	Stochastic gradient boosting and regression with Lasso regularization are used to improve police hiring decisions and teacher tenure decisions. The data used includes surveys as well as socio-demographic and classroom data.	Using machine learning models for hiring decisions can potentially reduce the excessive use of force by police and improve police-community relations. Similarly, students would benefit from better teacher hiring decisions.
Tax policy	Collaborative information acquisition for data-driven decisions (Kong and Saar-Tsechansky 2014)	Combinations of multiple learners and a variety of data sources are used to improve the cost efficiency of tax audit decisions.	The approach could increase sales tax profits by an average of 4 percent, strengthening this revenue source for governments.

Table 5: Examples of machine learning being applied to prediction problems from different areas of public policy making.

6 CHALLENGES AND LIMITATIONS

So far, this thesis has made the case for using machine learning in public policy making. Predictive modeling has the potential to change how public policy makers solve prediction problems. But machine learning is no panacea. This section explains what challenges and limitations machine predictions face. It is important for public policy makers to be aware of these challenges and limitations. In practice, most public policy makers will cooperate with professional vendors of prediction software or consultants rather than implementing a machine learning solution themselves. Companies offering machine learning solutions sometimes tend to oversell their products by making unrealistic promises (Kleinberg, Ludwig, and Mullainathan 2016). In such cases it is important to ask the right questions and not to take every claim for granted. This chapter aims at preparing public policy makers for this job. The challenges and limitations that machine predictions face in public policy making can be divided into three categories: the limits of prediction, technical and human challenges, and ethical and legal issues.

6.1 LIMITS OF PREDICTION

Can we predict everything? Of course not. Although we live in an interconnected world, where technical systems interact with the social nature of their users (Vespignani 2009) and where increasing amounts of data are produced by these interactions, some things remain unpredictable. Predictions of the behavior of complex techno-social systems of the kind that public policy makers face will never be completely accurate. The complex systems in today's world are comprised of large numbers of individual units, who exhibit to some extent random and unpredictable behavior (Brunner 1999). The world is constantly changing for all kinds of technical, environmental and social reasons. Because of this change, what was yesterday is not necessarily so today or tomorrow. Even given the biggest of all big data and the most

powerful algorithms, we are not able to predict the behavior of such systems with total certainty. Even machine learning cannot cross this fundamental border.

What does this mean in practice? When a system whose behavior we want to predict is too dynamic, a machine learning model that has been trained on data from the past, might not be able to accurately predict future data. This danger can never be completely avoided. In fact, one reason why complex techno-social systems are hard to predict is that policy makers constantly interfere with them. Take the hygiene prediction case study, for example. Given Yelp reviews and restaurant information, the machine learning model was able to predict hygiene violations in Seattle. If this model were used in practice, public health inspectors would be sent to the restaurants prioritized by the algorithm. Unfortunately, this very intervention might invalidate some of the model's future predictions. In the best case, this means that risky restaurants learn that they cannot escape hygiene inspectors anymore so that they improve hygiene conditions. In the worst case, restaurants learn that Yelp reviews can be revealing so that they incentivize their customers to write positive reviews or even buy fake reviews to conceal the opinion of their real customers (Kang et al. 2013). In general, there is evidence that big data and new prediction methods have not led to increased accuracy of the predictions of some reactive systems (Cueni and Frey 2014). Forecasting the weather is easy compared to predicting the behavior of systems that react to predictions. Unfortunately, public policy makers usually have to deal with the latter.

What can we do about the problem that predictive patterns in data change? The only solution is to update machine learning models from time to time. This is similar to what humans would do. Without the aid of a predictive model, hygiene inspectors have to choose themselves which restaurants to target. In some restaurants they find severe hygiene violations, in others they do not. This allows inspectors to learn from their experience where they most likely find hygiene violations. One could say that inspectors develop their own internal models to predict hygiene violations. Whenever they find no violation where they expected one, or when they unexpectedly encounter a violation, they should update their internal believes based on the new piece of

information. For using machine learning in practice, it is necessary to employ a similar strategy. Models have to be retrained from time to time to take newly-available data into account. This way it is possible to keep up with the changes in predictive patterns and to ensure that a model continues to make accurate predictions (Mackenzie 2015).

6.2 TECHNICAL AND HUMAN CHALLENGES

Data access, data management and computation are technical challenges in machine learning (Einav and Levin 2013). Creating a predictive model requires training data, but often public policy makers do not have access to the same wealth of data as the private sector. Because people are often concerned about the state's gathering too much data, accessing valuable big data for machine learning may pose a challenge for public policy makers. Even if policy makers get access to relevant big data, the next steps are not necessarily easier. Managing big data and computing with them requires substantial storage capacity and computational power. With datasets getting larger, even conceptually trivial task take longer and require more effort. For example, even such simple tasks as extracting and summarizing variables from big data and exploring the relationships between them can take considerable time (Einav and Levin 2013). Recruiting qualified personnel for such positions even poses a challenge to the usually well-paying private sector (Martin-Jung 2016). For the public sector, the availability of skilled personnel can be even more of a challenge (Hassani and Silva 2015).

Another technical challenge is to evaluate the accuracy of machine predictions. Decision makers want to see evidence that machine predictions are accurate before using them. In some areas of public policy making it is possible to run experiments. In the developing world, for example, randomized controlled trials are relatively common in education or public health and can be used to evaluate predictions. They are the best way to evaluate machine predictions (Kleinberg, Ludwig, and Mullainathan 2016). In other areas such as criminal justice, however, conducting experiments is impossible. In such cases, testing the accuracy of machine predictions is a big challenge. Take the bail prediction example again. As mentioned at the beginning of this thesis, the authors claim that using their machine learning algorithm,

"crime can be reduced by up to 24.8% with no changes in jailing rates, or jail populations can be reduced by 42% with no increases in crime rates" (Kleinberg et al. 2017). How do they know? They definitely did not run an experiment and jailed and released defendants based on their predictions. It is very difficult to evaluate what people who were kept in jail would have done if they had been released on bail. To overcome this problem, Kleinberg et al. (2017) had to use advanced econometric techniques exploiting the fact that defendants are as-good-as-randomly assigned to judges, some of which in turn make stricter bail decisions than others. The details of their econometric approach go beyond the scope of this thesis, but this example shows that evaluating machine predictions in the social realm can be very challenging. To be fair, evaluating human predictions is not necessarily easier. It is also very difficult to determine how well a single judge predicts how likely a defendant would commit a crime. In contrast to an algorithm that always predicts the same outcome given the same data, human predictions are also much more variable and often influenced by factors that should not play any role in decision making. There is evidence that judges are influenced by factors that should be extraneous to judicial decisions, including literally "what the judge ate for breakfast" (Danziger, Levav, and Avnaim-Pesso 2011). Using big data and properly evaluating predictions in such cases is a challenge, but also a chance to introduce more rigor to important areas of decision making in ways that were previously impossible (Bornstein 2017).

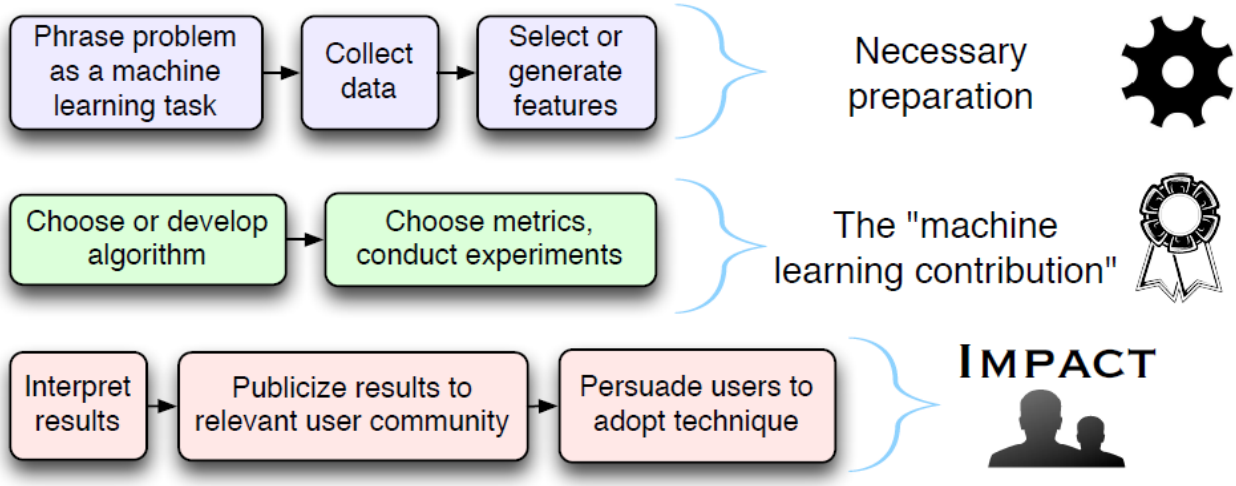

Figure 7: The long way from identifying a prediction problem suitable for machine learning to real-world impact (Wagstaff 2012).

Machine prediction are not only challenging in technical terms. There is a significant human component as well. Figure 7 shows that the way from identifying a prediction problem suitable for machine learning to data-based predictions having impact in the real world is very long. The steps between data collection and running experiments to evaluate machine predictions can roughly be regarded as technical steps. They mainly require the knowledge and expertise of machine learning professionals who have the right skills. For the steps at the beginning and the end of the way to real-world impact, however, the challenge is not so much technical but human. Humans are necessarily at both the beginning and the end of every implementation of a machine prediction solution. Everything begins with recognizing a prediction problem and phrasing it as a machine learning task. Public policy makers who have never heard of how machine predictions are possible likely have a hard time recognizing such opportunities at all (Porway 2015). There are many efforts to equip people with a more data-driven mindset, but such efforts are still in their infancy as data science is a very new discipline. If a public policy maker reads this thesis and only takes the point away that data and machine learning can be used for solving prediction problems, much is gained already.

The necessary step at the end of the way to real world impact is that humans actually use machine predictions for making decisions. Simply claiming that an algorithm makes accurate predictions is likely insufficient to convince people to use the predictions. Imagine a judge who has made bail decisions for all of her working life and suddenly is told that a machine makes better predictions than her. Convincing the judge to use the machine predictions in this case requires a good strategy. Such a strategy should not only teach that machine predictions are often superior to human predictions but also that machine predictions can be wrong. For example, hygiene inspectors have access to information that the machine learning algorithm in the case study did not know about. Somebody calling the hygiene authorities to complain about a problem in a restaurant gives human inspectors valuable information that the model based on online reviews cannot incorporate without further processing. The best possible solution to this problem is to combine machine predictions with human predictions so that the combination is better than either alone.

6.3 ETHICAL AND LEGAL ISSUES

Even if we can overcome the challenges mentioned so far, a crucial question remains: do we actually want to use machine predictions? Just because we are able to predict something, this does not mean that we are comfortable relying on that prediction (Kleinberg, Ludwig, and Mullainathan 2016). In this section we can only scratch the surface of the ethical and legal issues entailed by machine predictions, but public policy makers who plan to use machine learning should definitely be aware of their existence.

The biggest ethical and legal challenge that machine predictions face is bias. There are several ways how bias can enter machine predictions. The first way is the machine learning algorithm itself. It matters which machine learning algorithm is chosen to solve a prediction problem. Some algorithms are able to model more complex relationships than others. If a predictive model is too simple for the predictive relationships in the data, it underfits the data and introduces bias into its predictions. Only carefully testing what kind of machine learning algorithm is suitable for a prediction problem can avoid this issue.

The second way bias can enter machine predictions is by choosing the wrong prediction objective. The goal during the training of a machine learning algorithm is always to optimize some statistic of the training data. In the hygiene inspection case study, for example, the machine learning algorithm optimized for the area under the receiver operating characteristics curve (AUC). The model that generated the highest AUC was considered the best model. However, the metrics used to evaluate a machine learning model in the real world are often different from those used during the training of the machine learning algorithm (Lipton 2016). Hygiene predictions should be judged by how much they improve the allocation of hygiene inspectors to restaurants and improve food safety, instead of using such a complex metric as AUC that is obscure to everyone but machine learning experts (Wagstaff 2012). To align a model's prediction goal with the real-world outcome of interest, it is important that public policy makers sit down with machine learning experts to find an optimization

objective that reflects as far as possible the real metric of interest (The Economist 2016b).

The third way bias can enter predictions is through biased training data (Kleinberg, Ludwig, and Mullainathan 2016). This is the most difficult problem to avoid because biased data often implies that there is bias in the existing processes that generate the data. Imagine we wanted to predict how likely an individual is to commit a future crime. We gather data on arrests from the criminal justice system and other relevant information and train a machine learning model using this data. How could this be problematic? The problem is that the data from the criminal justice system merely tells us when somebody was *caught* committing a crime. But being caught committing a crime is an imperfect proxy for actually committing a crime. For example, if black people have been discriminated against by the police in the past, say by overly policing the areas where large parts of the black population live, and this discrimination is reflected in the data, machine learning models are very likely to pick up this discrimination and bias their predictions against black people (O'Neil 2016). In fact, there are already allegations that some of the crime prediction software in use in the U.S. is biased and makes worse predictions for blacks than for whites (Angwin et al. 2016). It is surprisingly difficult to keep bias out of a dataset. For reasons of justice, we might not want such information as race, religion or gender to play any role in predictions. It is easy to exclude such variables directly but very difficult to eliminate their indirect effects. The problem is that such data often enters predictions through the backdoor. Race, for example, is often strongly correlated with where somebody lives and gender with the profession somebody pursues. If we excluded every variable that correlates with a variable that should not play a role in prediction making from a dataset, there would often not be much left to make predictions on. The opposite of this problem, however, can also lead to bias. Instead of bias entering a dataset through the variables included in it, bias can also enter predictions through the variables that are not included in the dataset. It is hard to gather all relevant variables for a prediction, but predictions based on only some of all relevant variables can lead to omitted-variable bias and faulty conclusions (Kleinberg et al. 2017).

What to do against bias in machine predictions? Transparency, gathering more data and the continuous evaluation of machine predictions can help to some extent. Otherwise nobody would even know that, for example, some predictive models are more accurate for whites than for blacks. But even if a model has different accuracies for different subgroups of the population, is this reason enough to discard the model? Not necessarily, it always depends on the available alternatives to the machine predictions. If human predictions are even more biased or inaccurate than machine predictions, we might choose the machine predictions although they are not perfect. Probably neither human nor machine predictions will ever be completely free of bias, but being transparent about their shortcomings is the first step towards making them better. In general, public policy makers should be careful when using historically grown data for prediction making as parts of it might be biased against disadvantaged groups. There might be no way to completely avoid bias, but there are at least ways to mitigate the problem.

A second big ethical and legal challenge that machine predictions face, and which aggravates the bias problem, is that many machine learning algorithms are black boxes that do not explain the logic behind their predictions. Fully understanding a predictive model is possible only in some cases. The predictions of a linear model with few variables, for example, are easy to explain (Lipton 2016). The sums of the coefficients of the model multiplied by the values of the independent variables give the prediction. Based on the sign of a coefficient and its absolute value, it is easy to see how strongly any of the independent variables contributes to a prediction. The problem is that more advanced machine learning models almost always include non-linearities and complex dependencies between variables (Obermeyer and Emanuel 2016). In such cases, there is no single arithmetic formula to describe how a prediction is generated. Consequently, no human can understand such models in their entirety. This is obviously not ideal if we want to use such models for decision making. It is difficult to build trust in a model that we do not understand (Ribeiro, Singh, and Guestrin 2016).

What to do when a machine learning model is too complex to be understood by a human? There are two possible ways. One of them is to approximate the behavior of a model in the proximity of a single prediction we are interested in. Researchers have developed a technique that explains the predictions of any machine learning model in the region of the data around a prediction of interest by fitting a local interpretable model to this region (Ribeiro, Singh, and Guestrin 2016). This method does not explain the predictions of the model as a whole but at least gives an idea of how the model generates the prediction at hand. The other possible way is to use machine learning models that indicate how certain they are about a prediction. Some models do not only make predictions but also estimate the probability that a particular prediction is true. Of course, this does not allow us to explain the model's predictions, but at least we get an idea of how sure the model itself is about a certain prediction. Humans could then scrutinize those predictions in which the model only has low confidence. Maybe this is the only way to go if we want better predictions. The complexity of machine learning models might simply be the price we have to pay for high predictive accuracy (Welling 2015). For many real-world problems, there is a trade-off between the interpretability of a model and the accuracy of its predictions (Lipton 2016). Simple linear relationships might just not be sufficient to model the complex reality of our world. It is likely that many machine predictions are better than human predictions precisely because they model complex relationships beyond the boundaries of human understanding. As long is this complexity is counterbalanced by more accurate predictions, we might decide to use more complex models even if we cannot fully understand why a model predicts certain outcomes.

The last challenge in this section is the question of responsibility. Who is responsible for machine predictions? An answer to this question is relevant both from and ethical and a legal point of view. Some predictions such as in the bail example can have great influence on people's lives. If a prediction is correct, there is no problem, but just as human predictions, machine predictions are sometimes wrong. Who to hold accountable for a wrong prediction? The users of an algorithm? Those who created the model? Those who provided the data? There is no satisfactory legal answer to these questions yet. Existing legal systems are only slowly adapting to the new reality, but

first initiatives are on the way. One of the most far-reaching regulations in this area is the European Union's new General Data Protection Regulation, which gives people the right to meaningful information about the logic involved in automated decision-making that affects them (Courtland 2018). Along similar lines, the City of New York recently announced a task-force to scrutinize the algorithms used by the City for equity, fairness and accountability (City of New York 2018). These are good first steps to tackle this important issue, but some more have to follow before our public systems are fully ready to deal with machine predictions. Once the necessary laws have been adopted, machine predictions do not necessarily have to entail lower accountability than human predictions. People like to compare technological solutions to perfection, whereas humans are allowed to have their very human flaws (The Economist 2016a). But data-driven analyses and predictions can also be scrutinized for biases and other unwanted behavior. Mathematical models can even be more transparent and reliable than humans, who often offer mere justifications for their decisions instead of thorough causal explanations (Lipton 2016). Getting the legal basis for algorithmic accountability right is no easy task for a legal system that is used to deal with human agency, but there seems to be no way around this challenge.

This concludes the discussion of the challenges and limitations that machine predictions face in public policy making. There are some additional issues that should be taken into account in this context such as privacy, proper encryption and data protection (Lazer et al. 2009), which are definitely topics that play a role for machine learning because of all the data that is necessary to train a model. Security and gameability are other such issues that would merit some attention (Ghani 2016). However, as these issues are not unique to machine predictions but affect data handling and policy making systems in general, they are not discussed here. Nonetheless, a public policy maker who wants to use machine learning to solve a prediction problem should be aware of them in addition to the ones discussed in this chapter.

7 CONCLUSION

This thesis has shown that machine learning is a possible way to solve prediction problems in public policy making. Even though many challenges and limitations remain, data-based predictive modeling could change how public policy makers solve prediction problems. For the first time in history, large-scale big data on the social behavior of people is available. It is no longer statistical agencies with their coarse-grained, aggregated datasets that allow to gain the majority of insights into human behavior but the fine-grained data generated in today's techno-social systems (Davies 2017; Porway 2015). Living in such a data-driven world has many upsides for public policy making. Ideally, the focus on data could introduce more rigor and fairer processes. Humans have had to solve prediction problem in public policy making for centuries, but rarely ever have these predictions been thoroughly evaluated and tested for their accuracy. In the age of big data, this could finally change, with machine learning playing a major role in producing better predictions for the social good.

This being said, machine learning is no panacea. It should avoid being a tool in search of a problem. Public policy makers have many more problems to solve than just prediction problems, and some prediction problems cannot be solved by machine learning. Ideally, we combine data-based modeling with human intuition and experience to harness the strengths of both and avoid the weaknesses of either. It is unlikely that we will completely rely on algorithms for complex prediction problems such as bail predictions in the near future. But in many less controversial areas like hygiene violations this might well soon be the case. Unfortunately, there is also a great potential for the misuse of predictive modeling (Kleinberg, Ludwig, and Mullainathan 2016). Authoritarian states could use the technology to control their citizens. China, for example, plans to introduce a mandatory social credit system from 2020 on. Once the system is in place, the behavior of every citizen will be rated by algorithms. The resulting social score can then be used to "incentivize" behavior that is desired from the government's point of view (Botsman 2017). Combining this social credit system with machine learning could lead to a yet unknown degree of control of the government over its citizens. Instead of reacting to unwanted behavior after it has

happened, the government might opt for a preventive approach to stop unwanted behavior before the fact. Only the future will show whether machine predictions will be rather be employed for such use cases or for the ones outlined in this thesis. But machine learning and big data are here to stay – and public policy makers should be aware of the opportunities to use data-driven predictive modeling for the social good.

8 BIBLIOGRAPHY

Agrawal, Ajay. 2018. "The Economics of Artificial Intelligence." McKinsey Quarterly. 2018. https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-economics-of-artificial-intelligence.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2018. "A Simple Tool to Start Making Decisions with the Help of AI." *Harvard Business Review*, April 17, 2018. https://hbr.org/2018/04/a-simple-tool-to-start-making-decisions-with-the-help-of-ai.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias." ProPublica. 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Athey, Susan, and Guido Imbens. 2016. "The State of Applied Econometrics - Causality and Policy Evaluation." *ArXiv* 1607.00699. http://arxiv.org/abs/1607.00699.

Belloni, Alexandre, Victor Chernozhukov, Ivan Fernández-Val, and Christian Hansen. 2013. "Program Evaluation and Causal Inference with High-Dimensional Data." *ArXiv* 1311.2645. http://arxiv.org/abs/1311.2645.

Blumenstock, J., G. Cadamuro, and R. On. 2015. "Predicting Poverty and Wealth from Mobile Phone Metadata." *Science* 350 (6264). https://doi.org/10.1126/science.aac4420.

Booz Allen Hamilton. 2015. "Field Guide to Data Science." https://www.boozallen.com/s/insight/publication/field-guide-to-data-science.html.

Bornstein, Aaron M. 2017. "Are Algorithms Building the New Infrastructure of Racism?" *Nautilus*, December 21, 2017. http://nautil.us/issue/55/trust/are-algorithms-building-the-new-infrastructure-of-racism.

Botsman, Rachel. 2017. "Big Data Meets Big Brother as China Moves to Rate Its

Citizens." Wired. 2017. http://www.wired.co.uk/article/chinese-government-social-credit-score-privacy-invasion.

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures." *Statistical Science* 16 (3). https://doi.org/10.2307/2676681.

Brunner, Ronald D. 1999. "Predictions and Policy Decisions." *Technological Forecasting and Social Change* 62: 73–78. http://www.foehn.colorado.edu/nome/HARC/Readings/Brunner3.pdf.

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. 2016. "Productivity and Selection of Human Capital with Machine Learning." *American Economic Review* 106 (5): 124–27. https://doi.org/http://dx.doi.org/10.1257/aer.p20161029.

City of New York. 2018. "Mayor de Blasio Announces First-In- Nation Task Force To Examine Automated Decision Systems Used By The City." 2018. https://www1.nyc.gov/office-of-the-mayor/news/251-18/mayor-de-blasio-first-in-nation-task-force-examine-automated-decision-systems-used-by.

Courtland, Rachel. 2018. "Bias Detectives: The Researchers Striving to Make Algorithms Fair." *Nature* 558 (7710): 357–60. https://doi.org/10.1038/d41586-018-05469-3.

Cueni, Reto, and Bruno S. Frey. 2014. "Reaktivität Und Prognosen: Ein Irrtum Der 'Big-Data' Analyse." 2014. https://www.bsfrey.ch/articles/D_257_2014.pdf.

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso. 2011. "Extraneous Factors in Judicial Decisions." *Proceedings of the National Academy of Sciences* 108 (17): 6889–92. https://doi.org/10.1073/pnas.1018033108.

Davies, Wiliam. 2017. "How Statistics Lost Their Power." *The Guardian*, January 19, 2017. https://www.theguardian.com/politics/2017/jan/19/crisis-of-statistics-big-data-democracy.

Domingos, Pedro. 2012. "A Few Useful Things to Know about Machine Learning." *Communications of the ACM* 55 (10). https://doi.org/10.1145/2347736.2347755.

Einav, Liran, and Jonathan Levin. 2014. "Economics in the Age of Big Data." *Science* 346 (6210). https://doi.org/10.1126/science.1243089.

Einav, Liran, and Jonathan D. Levin. 2013. "The Data Revolution and Economic Analysis." *NBER Working Paper Series*. https://doi.org/10.1007/s13398-014-0173-7.2.

Gelman, Andrew. 2010. "Causality and Statistical Learning." *American Journal of Sociology* 117 (3): 955–66. https://doi.org/10.1086/662659.

Ghani, Rayid. 2016. "Keynote: Using Data Science for Social Good: Examples, Opportunities, and Challenges." https://www.youtube.com/watch?v=GTUozT9qxVw.

Glaeser, Edward L, Andrew Hillis, Scott Duke Kominers, and Michael Luca. 2016. "Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy." *American Economic Review* 106 (5): 114–18. https://doi.org/10.1257/aer.p20161027.

Hassani, Hossein, and Emmanuel Sirimal Silva. 2015. "Forecasting with Big Data : A Review." *Annals of Data Science* 2 (1): 5–19. https://link.springer.com/article/10.1007%2Fs40745-015-0029-9.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. *Springer Series in Statistics*. https://doi.org/10.1007/b94608.

James, Gareth, Witten Daniela, Hastie Trevor, and Tibshirani Robert. 2007. *An Introduction to Statistical Learning*. https://doi.org/10.1016/j.peva.2007.06.006.

Jean, Neal, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. 2016. "Combining Satellite Imagery and Machine Learning to Predict Poverty." *Science* 353 (6301): 790–94. https://doi.org/10.1126/science.aaf7894.

Jeong, Jig Han, Jonathan P. Resop, Nathaniel D. Mueller, David H. Fleisher, Kyungdahm Yun, Ethan E. Butler, Dennis J. Timlin, et al. 2016. "Random Forests for Global and Regional Crop Yield Predictions." *PLoS ONE* 11 (6): 1–15.

https://doi.org/10.1371/journal.pone.0156571.

Jordan, M. I., and T. M. Mitchell. 2015. "Machine Learning:Trends, Perspectives, and Prospects." *Science* 349 (6245): 255–60. https://doi.org/10.1126/science.aaa8415.

Kang, Jun Seok, Polina Kuznetsova, Stony Brook, and Michael Luca. 2013. "Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews." In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1443–1448. http://www.aclweb.org/anthology/D13-1150.

Kleinberg, Jon, Jens Ludwig, Molly Cohen, Alexander Crohn, Gretchen Ruth Cusick, Tim Dierks, John Donohue, et al. 2017. "Human Decisions and Machine Predictions." *NBER Working Paper*. https://doi.org/10.3386/w23180.

Kleinberg, Jon, Jens Ludwig, and Sendhil Mullainathan. 2016. "A Guide to Solving Social Problems with Machine Learning." *Harvard Business Review*, December 8, 2016. https://hbr.org/2016/12/a-guide-to-solving-social-problems-with-machine-learning.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review: Papers & Proceedings* 105 (5): 491–495. https://doi.org/10.1257/aer.p20151023.

Kong, Danxia, and Maytal Saar-Tsechansky. 2014. "Collaborative Information Acquisition for Data-Driven Decisions." *Machine Learning* 95 (1): 71–86. https://doi.org/10.1007/s10994-013-5424-x.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, et al. 2009. "Computational Social Science." *Science* 323. https://doi.org/10.1145/2556420.2556849.

Li, Zhidong, Bang Zhang, Yang Wang, Fang Chen, Ronnie Taib, Vicky Whiffin, and Yi Wang. 2014. "Water Pipe Condition Assessment: A Hierarchical Beta Process Approach for Sparse Incident Data." *Machine Learning* 95 (1). https://doi.org/10.1007/s10994-013-5386-z.

Lipton, Zachary C. 2016. "The Mythos of Model Interpretability." *ArXiv* 1606.03490. http://arxiv.org/abs/1606.03490.

Mackenzie, a. 2015. "The Production of Prediction: What Does Machine Learning Want?" *European Journal of Cultural Studies* 18 (4–5): 429–45. https://doi.org/10.1177/1367549415577384.

Martin-Jung, Helmut. 2016. "Edel-Informatiker Gesucht." *Süddeutsche Zeitung*, March 19, 2016. http://www.sueddeutsche.de/karriere/mangel-an-it-experten-auf-der-suche-nach-dem-edel-informatiker-1.2908223.

Mol, Christine De, Eric Gautier, Domenico Giannone, Jeffrey Wooldridge, Sendhil Mullainathan, Lucrezia Reichlin, and Herman van Dijk. 2015. "Big Data in Economics: Evolution or Revolution?" In *Economics without Borders*. http://www.coeure-book.ceu.edu/CH14_BigDataMethods_Feb8.pdf.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106. https://doi.org/10.1257/jep.31.2.87.

Munoz, Andres. 2014. "Machine Learning and Optimization." 2014. https://www.cims.nyu.edu/~munoz/files/ml_optimization.pdf.

O'Neil, Cathy. 2016. *Weapons of Math Destruction*. https://weaponsofmathdestructionbook.com/.

Obermeyer, Ziad, and Ezekiel J. Emanuel. 2016. "Predicting the Future — Big Data, Machine Learning, and Clinical Medicine." *N Engl J Med* 375 (13): 1216–19. https://doi.org/10.1056/NEJMp1606181.Predicting.

Porway, Jake. 2015. "Five Principles for Applying Data Science for Social Good." 2015. https://www.oreilly.com/ideas/five-principles-for-applying-data-science-for-social-good.

Potash, Eric, Joe Brew, Alexander Loewi, Subhabrata Majumdar, Andrew Reece, Joe Walsh, Eric Rozier, et al. 2015. "Predictive Modeling for Public Health: Preventing Childhood Lead Poisoning." *KDD '15- Proceedings of the 21th ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining, 2039–47. https://doi.org/10.1145/2783258.2788629.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." *ArXiv* 1602.04938. http://arxiv.org/abs/1602.04938.

Rudin, Cynthia, and Kiri L. Wagstaff. 2014. "Machine Learning for Science and Society." *Machine Learning* 95 (1): 1–9. https://doi.org/10.1007/s10994-013-5425-9.

Schroeder, Ralph. 2014. "Big Data: Towards a More Scientific Social Science and Humanities?" In *Society and the Internet: How Networks of Information and Communication Are Changing Our Lives*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199661992.003.0011.

Taylor, L., R. Schroeder, and E. Meyer. 2014. "Emerging Practices and Perspectives on Big Data Analysis in Economics: Bigger and Better or More of the Same?" *Big Data & Society* 1 (2). https://doi.org/10.1177/2053951714536877.

The Economist. 2016a. "Frankenstein ' s Paperclips." *The Economist*, June 25, 2016. https://www.economist.com/special-report/2016/06/25/frankensteins-paperclips.

———. 2016b. "Of Prediction and Policy." *The Economist*, August 20, 2016. http://www.economist.com/news/finance-and-economics/21705329-governments-have-much-gain-applying-algorithms-public-policy.

U.S. Department of Justice. 2016. "Persons Arrested." Crime in the United States. 2016. https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/topic-pages/persons-arrested.

Varian, Hal R. 2010. "Computer Mediated Transactions." *American Economic Review* 100 (2). https://doi.org/10.1257/aer.100.2.1.

———. 2016. "Causal Inference in Economics and Marketing." *PNAS* 113 (27). https://doi.org/10.1073/pnas.1510479113.

Vespignani, A. 2009. "Predicting the Behaviour of Techno-Social Systems." *Science* 325: 425–28. https://doi.org/10.1126/science.

Wagstaff, Kiri. 2012. "Machine Learning That Matters." *Proceedings of the 29th International Conference on Machine Learning*. https://doi.org/10.1023/A:1007601113994.

Welling, Max. 2015. "Are ML and Statistics Complementary?" In *Roundtable Discussion at the 6th IMS-ISBA Meeting on "Data Science in the next 50 Years."* https://www.ics.uci.edu/~welling/publications/papers/WhyMLneedsStatistics.pdf.

Wu, Jionjlin, Jason Roy, and Walter F. Stewart. 2010. "Prediction Modeling Using EHR Data: Challenges, Strategies, and a Comparison of Machine Learning Approaches." *Medical Care* 48 (6): 106-S113. https://doi.org/10.1097/MLR.0b013e3181de9e17.

9 APPENDIX – PREDICTION FOR CAUSAL INFERENCE

Machine predictions alone are not sufficient to answer causal questions because they only predict that something happens, but not why. However, machine learning can still lead to more robust causal inference. Predictions are an essential part of the effects-of-causes framework of causal inference. As the name suggests, in this framework we are interested in what the effects of a cause are. We can do this in the common counterfactual formulation. What would happen if we did Y instead of X? By definition, a counterfactual is never known – but we can predict it. Consider the four popular techniques for identifying the effects of causes: randomized controlled trials, difference-in-differences, regression discontinuities and instrumental variables. For those who are familiar with these methods, here are some points on how to use machine learning to improve causal inference (Varian 2016):

- In randomized controlled trials we need to estimate the counterfactual: what would have happened to the treated without the treatment and to the untreated with the treatment? Machine learning models can do this job, taking a large number of variables into account. This also makes it easier to estimate heterogenous treatment effects (Belloni et al. 2013).
- The same is valid for difference-in-differences, for which we have to estimate the counterfactual that would have happened without the intervention that took place. Machine learning's excellent prediction abilities might help in this respect.
- For regression discontinuities we need a predictive model for the behavior near the discontinuity. Given enough data, this behavior can be modelled nonlinearly with machine learning.
- In the case of instrumental variables, the instrument should enter the equation linearly, but the other covariates do not have to. Machine learning can be used to model flexible functions of these covariates.

All of these tools are used for causal inference, but machine learning and big data may be able to improve these tools, giving us more trust in the causal estimates derived from them.