Project Summary

Predict Customer purchase behavior - Product Return rate

Nowadays machine learning modelling techniques become more and more popular, mostly because of its applicability to our everyday life. Talking about retailers, data-driven solutions applied after proper understanding of "causality effects" are major defining elements of the business success. Machine Learning refers to a capability of a certain program to auto-improve, or learn from data, in order to show to business owners important insights, solve issues, and create additional value benefits.

The main objective of this project was to implement Machine Learning power, in order to predict the customer return rate for an American ecommerce company. This non-trivial task is almost impossible to solve for human mind accurately, especially if dealing with big data sets and considering somewhat limited data knowledge. Despite this, mathematics in a combination with statistics, can easily make a proper guidance for businesses. We implemented several variations of a Random Forest Machine Learning Model to a given data set, and predicted weather the customer will make a return of goods, or not.

The target of the project was to analyze the return rate, by feeding the data to an algorithm (Random Forest in this case). In order to skip the complexity of this algorithm, we'll try to build an example of the current case study, so everybody can understand it clearly. The available data set consists of 10000 records of historical sales of a retail store. Each data record indicates the Order Date (day, month, year), Dispatch Date, Shipping Type, specific customer information, Segment, Country-State-City info, Product related data (like product category, sub-category, quantity or products ordered), Price, and weather discount was applied or not. The Random Forest algorithm aims in our case to detect if a particular order will be returned, or not. The stuff inside an algorithm, builds a tree-based structure, based on existing data features. Features in this case are characteristics of each data point, meaning:

Price, Quantity, Segment, Category (mentioned above). This problem is so-called 'Supervised classification problem', it means that we have historical data set, that already has inside target feature inside (return value, 'yes' or 'no', in this case). Supervised – because we feed this data into algorithm, and it tries to learn from the past data, in order to make a prediction later on the unseen data. Classification – because we want to classify our new data into 2 classes: 1) Order returned; 2) No order return. To continue, the Random Forest algorithm learns from our data, and stores learnt parameters inside, will be the order returned based on given feature combinations.

In our case we were dealing the problem called 'Imbalanced classification problem'. Imbalanced classification occurs, when the vast majority of data points (90 per cent or more) belongs to a certain class. In this project, the return orders occupied approximately 5 per cent of the data, which is extremely small portion. That is why, when dealing with Imbalanced classification, we need a lot of data, in order to be able to give the algorithm all possible combinations of 'why this order was returned?'. Given data set didn't have much data, however, the obtained results are quite promising.

After the analysis, we obtained a model which can make a prediction if a customer will return a good after buying, or no. The Precision of this mode is 97 per cent, which is stands for a high value. As the data set consists of predominantly orders without return, this measure can't be a standalone characteristic of a model as such. That is why we introduced 3 other measures of model accuracy: Sensitivity, Specificity, and Balanced Accuracy. Specificity stands for correct measurements of major class (non-return status), while sensitivity measures how good is model for minor class (orders with return status). Balanced Accuracy, hence, is a generalized performance of the model, this metric can be used in order to judge the model's work. Sensitivity score of the final model is 70 per cent, while Specificity score is 76 per cent, with Balanced Accuracy of 73 per cent. So, to simplify, we can say, that our model is accurate in 73 cases out of 100. The question is, is it good or bad, and why do we need to calculate specifically Balanced Accuracy? Well, let's try to explain in simple terms. Imagine that you have 100 orders in total: 95 orders without returns, and 5 order with returns; imagine, that you know that out of this

hundred orders, 5 orders have return status, and you need to make a guess, which of them. Well, you can make a simple random guess, but let's say you decided to make a guess, that none of orders have return status. You may say that it's wrong, and you'll be 100 per cent correct. But wait, let's calculate the model accuracy. Even you made a mistake in 5 points, you were correct in 95 point. Furthermore, your accuracy score will be 95 per cent. Seems high, right? But we all know, that you completely mislead return orders. The main idea of this approach, is to maintain 'the golden mean' and to be able to predict both classes, even with some mistakes. Moreover, the data set used in this project is relatively small, and still we were able to achieve this acceptable accuracy of 73 per cent! Imagine, how the performance will increase, when you have more data. The beauty of machine learning models, is that they are improving by re-training itself with time, when more data become available.

In order to understand, why customers making a return, we can checked features Importance after modelling. We extracted information from our Machine learning model, and examined the results. Below you can see top 7 features, which have an influence on Returns:

Region_West	0.69866286
Year_2014	0.39560446
Gap_0	0.39309598
Gap_6	0.38641486
Region_East	0.36623886
Segment_Corporate	0.35445843

In the table above, left column indicate a Feature name (where: Region West – orders from West Region; Year_2014 – orders done in 2014; Gap_0 – delivery was done on the same day as order; Gap_6 – it took 6 days to deliver an order; Region East – orders from East Region; Segment_Corporate – orders of corporate segment), and right column indicate feature importance – 1 – the strongest one, 0 – no relation, weak importance.We clearly see, that people tend to return goods more, when they are from the West. Moreover, we can conclude, that fast

(gap_0) and long (gap_6) delivery can trigger order return as well. People from Eastern region also tend to make returns, but with lower frequency.

First of all, let's talk about delivery time, more specifically about Fast Delivery (Gap_0) and Long Delivery (Gap_6): from the results above, it may be seen, that when delivery time is too fast (same day delivery), people who order a particular good realized, that they don't need any more this item, as a result return option was triggered. On the other hand, when delivery time is long, like 6 days, maybe person who made an order expected to get this item faster (2-3 days), or even maybe it was something urgent. In this case, person could buy the same good at a different store, and after he/she received an order, it was not needed any more, so return option was triggered.

In case of Western Region orders, we concluded, that most of Returned Order came from this part of the country. We can assume, that it took too long, in order to get an ordered item, without knowing a prior location of you store. We recommended to work with these customers to expedite the delivery of the product, and maybe offer a small discount, if model shows that a person is going to produce a Return.

To conclude, this project showed proof of concept of applicability machine learning power, in order to find unsatisfied customers with returns. Current issue with unsatisfied customers were turned into advantage with ability to predict which customer is most likely to return the order. By knowing this information, the company can easily adapt their business to customer needs, by applying, let's say offering extra discount to the customer likely to return the product, thus ensuring they don't return the goods. In this case company's market will grow, and they will not lose customers. This is extremely useful when it comes to marketing. Moreover, the company can plan campaigns ahead, with predefined knowledge of a return rate.

The value of machine learning, and the reason why retailers of all sizes need to start exploring it, ultimately lies in deriving more insights from the available data. In the end, Machine Learning can provide a major bottom-line lift to the overall business.