## Public Project Summary

#### Business objectives:

The primary aim of my project was to improve a questionnaire-based recommendation algorithm. I did this for a startup company called BrokerChooser which is part of the CEU iLab. The final aim was not fully accomplished, since it turned out that the data collected by the company does not make it possible. Despite this I squeezed out most of the business value based on the available data and gave advice on how to conduct the questionnaire and collect data to optimize the recommendation algorithm in the future. While working on this project I learned a lot including:

- How to use R more thoroughly
- How to make a dashboard with Shiny in R (I used widgets, modified data, etc.)
- How to give descriptive statistics for non-technical people (mainly using ggplot2)
- How to do PCA based on a questionnaire
- How hard it is to share my data related problem with non-technical people
- How much time each data related task required in real life.

BrokerChooser has a **questionnaire-based broker recommendation system** which works in the following way: They ask 8 questions on their website and based on the answers given they show an ordered list of 16 brokers from the most suitable to less suitable brokers. Then the users can compare the brokers and pick one which they think is the best suited to their needs. Based on the terminology of the company a broker means a trading platform which makes it possible to trade financial asset via your computer or your mobile phone. They collect each answer given to this questionnaire and the broker chosen. Data is also collected about which brokers were recommended for a specific answer combination.

Besides the algorithm optimization task other tasks came up during the project that I consider to be a main task as well. The two most important tasks were the following:

- 1) **Give descriptive statistics** about how people filled out the questionnaire to find out the characteristics of these people visiting the website of the company. With the help of this analysis the company can get insights about the behaviour of their customers and get ideas how the questionnaire can be improved. (e.g.: whether a question is important or not, specific target groups)
- 2) Give advice about how the data collection method should be changed and how to optimize the broker recommendation algorithm in the future.

## The questionnaire and how the broker recommendation system works:

## The questionnaire consisted the following questions and answers:

Most of the questions are radio button ones, which means you can select only one answer. Except the 7<sup>th</sup> question where you can select multiple answers.

- 1) Choose which one are you: a) starter, b) investor, c) trader
- 2) How independent are you? a) I want a professional to manage my money., b) I might need to consult a broker every now and then., c) I am interested in following other trader's strategy (social trading)., d) I make the decisions myself on where to put my money.
- How much trading experience do you have? a) I am a beginner (e.g. had only funds so far), b) I know the basics (e.g. already traded with stocks), c) I can deal with more complex transactions (e.g. placing a stop-loss order), d) I am a professional (e.g. trading with options, futures)
- 4) How much money would you invest? a) less than 10.000 USD, b) more than 10.000 USD
- 5) How often would you place an order? a) I trade on a daily basis, b) I keep an eye on the markets and trade every few weeks, c) Buy and hold: after buying a stock I would hold it in the medium term
- 6) Which markets would you like to reach? a) It's ok to have access only to the major markets (e.g. USA, Japan, Hong Kong), b) I want to reach the smaller markets as well (e.g. South Africa, Turkey)
- 7) In which products would you like to invest? (you can select multiple items) a) simple investment products (e.g. funds, stocks) b) more sophisticated investment products (e.g. corporate bonds), c) leveraged products with higher risk (e.g. CFD, options), d) FX trade
- 8) What type of investor protection do you want your broker's country to provide? a) I need investor protection, but riskier countries are fine as well (e.g. Cyprus, Malta), b) I want the investor protection of a country with minimum AA credit rating (e.g. USA, Switzerland)

#### The recommendations system of the company works in the following way:

At the 3<sup>rd</sup> question the user selects how much trading experience they have. Based on this answer the company has a predefined ordered broker recommendation list. These predefined lists were created by domain knowledge and intuition. Based on every answer given during the questionnaire the user excludes specific brokers from the ordered

predefined list resulting in a subset of the list. This subset is going to be the final broker list for a given answer combination.

## The project outcome:

Since the implemented method of data collection has not made it possible to do complex modelling tasks I ended up creating a dashboard. I believe that combining the results displayed on the dashboard with domain knowledge of the decisionmakers in the company can lead to valuable findings. On the dashboard they can compare KPIs and plots of different outcomes to each other. See what the dashboard can display in the next chapter. I think this dashboard is a useful tool for making business decisions, optimize the broker recommendations and to show the problem regarding the currently used data collection method.

The other important part of the project was to give advice regarding how to change the data collection from a data analysis perspective to be able to optimize the algorithm in the future. I summarise the main points of the problems and my advices given in the following lines:

- To suggest more optimal algorithms, we would need different recommendation algorithms to race with each other to determine which one performs better. Currently the company has only one recommendation algorithm which works very deterministically. We cannot see the effect of different recommendation systems to decide which one is the best. Creating multiple recommendation algorithms and set a KPI for them to race would be an ideal way to move towards the most optimal algorithm.
- Summarizing the point above: There is no way to conduct A/B tests based on the current data to decide what would happen if we had given different kinds of recommendations to the users.
- There is no user ID in the data, so we cannot distinguish those who filled out the questionnaire and clicked multiple brokers at the same session. Therefore, I suggested that the company implement a measure to be able to collect this metric.
- There is no timestamp in the data. This would be useful as we might recognise time related patterns and ease data filtering in the future.
- Add a variable to the data indicating the **version of the algorithm** and the **version of the questionnaire** to track modifications.
- Now we base our decision on click on brokers rather than account openings. This can be a false measure.
- There are default answers based on the 1<sup>st</sup> question. There is no way to measure its effect. The company should measure the effect of the default answers.

## The Dashboard:

The dashboard is based on data imported from Google Analytics. I chose a 5-month period where there were no modifications of the survey and the broker recommendation list. After several transformations I could put together an easily analysable data table structure. I transformed the original data from wide to long, created individual observations from aggregated ones, renamed some variables and cleaned the data. I ended up with around 7000 observations. Out of the 7000 there were around 2000 observations who not just filled out the questionnaire but chose a broker at the end of the questionnaire process.

#### How the dashboard looks like:

The dashboard can be found on the following link: <u>https://krinya.shinyapps.io/dashboard\_upload/</u>

#### The dashboard can display the followings:

- The number of observations
- The most common answer combinations this can be obtained by the fill\_id-s from the "Top 20 most common IDs plot"
- Which broker do people choose and in which proportion this can be seen from the "Broker distribution" plot
- What the recommended brokers were for every ID and which brokers were recommended most frequently at the 1<sup>st</sup>, 2<sup>nd</sup> and 3<sup>rd</sup> place of the broker recommendation list.
- How the answers were given and in which proportion.

# The dashboard consists of two columns. The functions of the two columns are completely the same. These two serve comparability. This way plots can be easily compared to each other.

You can filter the data for a specific subset at the top of the dashboard under the "Questionnaire properties" text. By doing this you can filter for a group to look after a phenomenon manually. You can filter for the followings:

- Choosing a broker at the end of the questionnaire or not
- Fill ID (even multiple ones) fill\_id is a number used internally by the company to identify an answer combination given to the questionnaire

- Broker name to display data about only those who ended up selecting a specific broker at the end of the questionnaire
- The questions themselves. e.g: if you want to filter for those who stated that they want to invest more than 10,000 \$ you can do that under the "How much money?" tab.

#### A simple example from the dashboard:

If we filter for those who would like to invest less than 10,000\$ on the left side and filter for those who wants to invest more than 10,000\$ on the right, we get the following results displayed on a plot of the "How much trading experience do you have" question:



Screenshot from the dashboard

We can see that those who want to trade with less than 10,000\$ are less experienced. (Or at least that is what they say about themselves)

#### Other things I did:

I tried to use machine learning to predict which broker a person would choose based on the currently available data. I beat the recommendation algorithm of BrokerChooser by a small margin, but I suggested not to implement this method. Instead of this I suggested a new method to change the data collection method and to optimize the broker recommendation algorithm in the future. In short, I proposed that they implement many kinds of recommendation algorithms and race them based on a selected KPI such as the number of clicks. I also encourage them to use random recommendations for a small subset of the visitors to set the baseline of the effect of the recommendations.

Additionally, I did PCA on the questionnaire to see which questions and answers are important.

#### Conclusion:

The primary goal to optimize the questionnaire-based broker recommendation algorithm of the company BrokerChooser was not fully accomplished since the conduction of the data collection in the past has not made this task possible.

After realising the problem, I set other goals which I believe created business value for the host company and fulfilled the requirements of the capstone project. I did the following things:

- Made descriptive statistics based on the current database
- Created an XGBoost model which beats the currently used broker recommendation algorithm
- Created a dashboard to display some plots and KPIs regarding how the users fill out the questionnaire, which broker they choose and which brokers are being recommended for them
- Recommend a better approach how the data should be collected ideally and how to optimize the broker recommendation algorithm in the future

In hindsight I would choose a different project because I really wanted to use machine learning models. I feel like I could do much more in that field where proper data is available. Although the project was not ideal it made me realize how real live situations differ from those in a prefect data environment. Especially when it comes to young start-ups where data collection methods are not the most sophisticated.

I believe that if the methods I recommend are implemented the project can yield great results.