Public Project Summary fo Capstone project

The scope of the project was to put together a proof of concept for retail banks, to help in their financial target allocation of banking products. The project was based on a previous initiative, that had some lacklusters. After carefully understanding the previous project's outcome, I tried to incorporate every usefull piece in the proof on concept I wanted to create, and offer a solution that is the closest for the previous client's needs.

After a consultation with an external expert, I ensured myself, that there is no estabilished methodology for demand prediction upon doing estimation through instable peers. Therefore, I decided that the proof of concept project will have the following goals:

I. helps in careful feature selection regarding target allocation based on the original form of the data, without assigning proportional scores, or relative intuitive wieghts to them
II. helps in resource planning: the model should include branch level bank data
III. grants insight on the current performance of the branches

Together with the external consultant who assigned me to the project, we have made the decision, that as the target variable, I should re-estimate the real performance of banking products of which it is not possible to estimate raw market demand.

The proposed system uses the original inputs of the bank, and the microlocation level data collected by the company I am working for. The modeling happened in R, and the visualization happened using Tableau.

The key takeaway while defining the scope was, that sometimes the consultancy business meets with unmatcheable client expectanceies, and it is hard to both provide a scientficly correct solution, and fits a budget. Even while just internally rewamping the project, I met several times the comment: „you are not rigth abot the solution being incorrect, it met the client's needs, and it has been paid, I don't even know why you

still care about that". In my perspective, currently consultancy businesses (at least in big4 companies) do not utilize the data assets given properly, and it is hard to interpret something more complex than an excel sheet.

After successfully defining the scope of the project, I started the data collection from the previous project. Cleaning the data is alwasy a demanding task. The main challenge was the format here: overformatted excel tables are not compatible with statistical softwares. Data cleaning was must struggle. After consulting with an subject matter expert, I selected the features, and cleaned the data. The result ended up being a „wide" dataset, with not too many rows, but a lot of feature variables.

When the data was ready, the first decision I had to make was to use time series model or cross-section model. Essentially, the problem statement would suggest time series analysis, to predict next year's possible sales number for each product, however during data collection research, I realized, that there is no available yearly data on microlocation level, only just for a few parameters. Therefore I decided, that I have train a cross section model, and performance analysis will happen based on a re-prediction of the existing sales data in the cross-section.

For the analysis I will separate the dataset into two parts: the training (70% of the data) and the test set (30% of the data). I will train the algorithms on the training set, and will run them on the test set. I will chose between the two algorithms based on the Residual Mean Squared Error calculated on the test set. In this case, when I created the test and the data set, I made sure, that I will divide according their proportions the Budapest brancehs (70% of the Budapest branches are assigned to the training, 30% of the Budapest branches are assigned test set). In a normal scenario random sampling would be beneficitial, but with this low amount of data rows, I wanted to make sure, that neither the training or the test set is disorted based on that feature.

After the train and the test set was created, I had to decide which algorithm I should use. For wide datasets, I had to decide to use algorithms, that are robust to noise, and perform a careful fature selection. Therefore, I have chosen elastic net, with both boostrapping and crossvalidation methods, and extreme gradient boosting to compare. While using elastic net, the algorithm that I have known from my studies

with both setups, it was exciting to see to receive the expected differencences from both variation. I expected to receive more endogenous, but fewer variables, and more numerous, but less likely endogenious variables from crossvalidation methodogoly. Learning extreme gradient boosting was challenging, new, exciting experience. Testing both algorithms on the training set I received no significant difference, in some cases one or other performed better. Since interpretation of the coefficients had a high priority during this project, I decided to use elastic net for the final predictions.

To properly interpret the model, I decided to use Tableau for data visualization. These visulaizations are focusing on the representation of the findings of the underlying analysis.

The visulaziation was created in the form of 3 tableau dashboards for each target product.

All 3 dashboard have 4 main elements:

I. Map, that indicates performance of the banking product:
- For each branch, a circle is allocated based on GPS coordinates. The bigger the circle is, the bigger the absolute difference is between the predicted and the actual value. The color indicates the relative difference between the predicted value and the actual value, compared to the original value. Red indicates negative difference, blue indicates positive difference. relative and absolute differences are added as a calculated column in tableau.
- The map has two actions:
    - Hover: hovering on the circle of a branch will visualize all the properties of the branch
    - Select: selecting branch(es) will cause their name to reveal on the map. Only the selected branches will be present in the two bar charts below the map

II. 2 side-by-side column charts: the top side-by-side chart shows the 3 most important local feature that has a positive effect on sales by the name of the branch, the bottom one shows the 3 most important local feature that

has negative effect on sales. By default it shows all the branches, selection on the map effects both side-by-die column charts

III. imported picture from R showing the feature importance for each product

Currently there is two other action on the dashboard that can be taken: filter by relative performance using the slide, and a manual selector for brancehs.

The proof of concept system I created selected the most important features for 3 core banking producst. During modeling, I have succesfully trained the algorithms to show partial effects of both branch and local properties. With the help of tableau visualization, I also granted insight how well branches are performing.

However providing descriptive analytics and key findings regarding the product sales would be insightful, it has no real business value in this scenario, therefore it is out of the scope of the project.

The inmediate use of this proof of concept is quite obvious: if the given bank would open a new branch anywhere in the country, my company would easily provide consultancy for them, both resource and financial product target allocation could happen inmediately.

The next step would be to fully automate the solution, that only requires minor finetuning. The only input it would require is the data table from a new bank containing their branch level data

The further improvement of the project would be to sell itt o an another bank, ith would strenghten our inner database with more sales data, and the algorithm could be retrained with more data points. Also it would be helpful, to channel in their own key painpoints, and further develop both the code and the dashboard.

In general, I can conclude, that data based consultancy is not just the future, but the present. Even this capstone project offers more insightful analysis for the retail banking client, than the previous solution that was created under 6 months by 4 people. Firms, who do not invest in data science capabilities will slowly lose their market potential.