AALTO UNIVERSITY

MS-E2177
Seminar on Case Studies in Operations Research

Interim Report

Otto Saikkonen

Oscar Björklund

Aki Malinen

Iiro Sallinen

Juho Heimonen

June 10, 2020

1 Project Status and accomplished tasks

Our project was initially divided in 4 work packages:

- Project planning and background research
- Project initiation
- Technical execution
- Project management

The first part was finished on schedule in March, as was the project initiation phase - we received an initial description of the data structures from Nordea with a slight delay, but this was included in the previous project plan report. The start of the technical execution was delayed by about two weeks compared to our initial project plan bringing the beginning of the technical execution close to the exam period where no activities were planned. This longer delay was due to client side difficulties in authorizing the usage of, querying, compiling, encrypting and transferring the data taking into account the required security measures on the client side. Further, the outbreak of COVID-19 in Finland posed a challenge for the communication with the client, who saw their workload rise due to the outbreak.

We received data from three different data sources, which was a small subset of the data promised to us following our first meeting. Most of the work we had done, based on the database's schema we received earlier, was now all for nothing due to the missing attributes in the data sources. There was no way for us to calculate CLV according to our original plans. In addition, we had some problems reading the data, since the 27 zip files sent to us were somewhat corrupted. We were able to fully extract the product data and 6 semi-disjointed months of customer time series data.

In our preliminary data analysis we noticed that 8 out of 9 features in the time series data were zeros for all the customers in 4 of the 6 months we had (Figure 3). The product data has all the customer's purchases, and it also includes the customer's age at the time. We hypothesized that customer's age is a crucial feature for our project. However, during our preliminary data analysis, it came clear that customers' ages were altered, most likely due to privacy reasons. Since originally the data manipulation was probably not done in very sophisticated way (Figure 4), some of the information it held is most likely lost in the process.

Due to the complications with the data, we decided to momentarily set aside the CLV model and focus on identifying key properties of Nordea Life Insurance customers until further information is received. This was done by building a model that estimated the probability of customer owning a MyLife product. Given the time resources available for this purpose, the model was a simple gradient boosting classifier, which actually performed relatively well. The area under ROC curve (AUROC) (Figure 5), precision and recall of our classifier were 0.91, 0.99 and 0.58, respectively. Moreover, we were interested in those features which would be important in determining whether customer has MyLife products or not. Figure 6 describes the feature importances extracted from our model. It is clear that other Nordea products play a key role here. This is something we will continue to investigate further on.

After we had cleaned and familiarized ourselves with the data, we started implementing a first version of a CLV model using Python. The first CLV model of new customers is now complete. Based on the following parameters: gender, risk grade, age, income, savings and what other Nordea products the customer has. Some important factors, such as churn rates, are still not included in the model. The model still has to be expanded, validated and trained before we can present preliminary results. Overall, we are almost on schedule and the project is progressing continuously.

The scope and objectives of the project were not set on stone by Nordea at the start of the project. The objectives were negotiated with room for interpretation. Since the workload and feasibility of the defined objectives has been reasonable, no changes have been made to the objectives and scope of the project. However, if project timeline turns out to be harder to achieve than expected, less resources can be put to implementing the results into concrete business implications as this part was seen secondary as it requires more insider knowledge of the existing internal strategy.

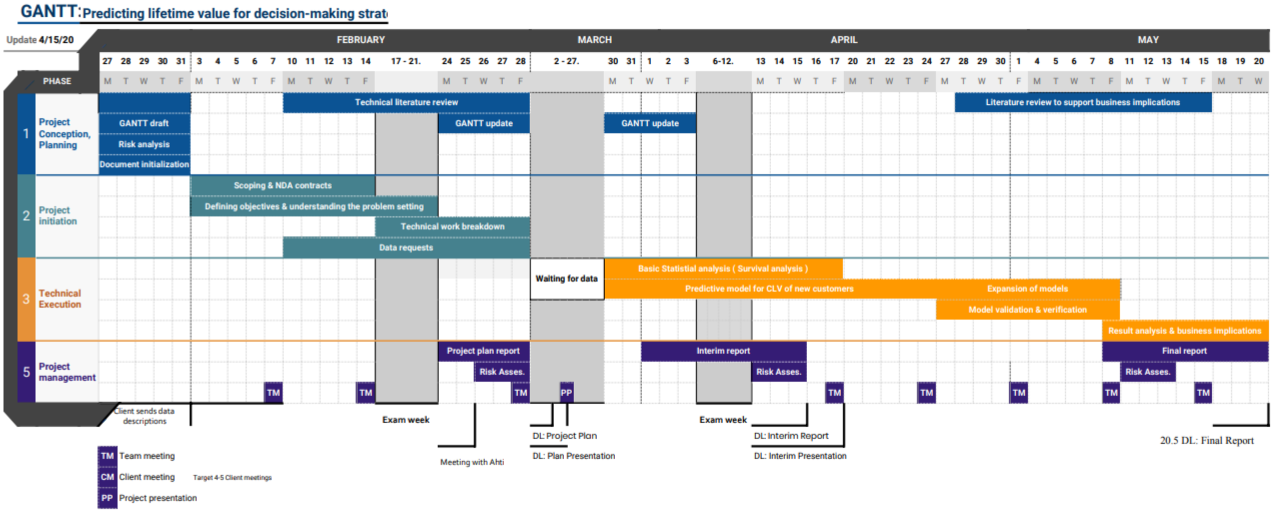

Figure 1: Project GANTT. Larger version available in the Appendix 1

2 Project Plan for the remainder of the project

These work package divisions were not created based on deliverable items, in contract to a common practise in project management. Deliverable based division is usually relevant when the deliverables are the main outcomes. In the case of this project, we see that the course deliverables (reports) are **not** the main outcome of the project, instead it is the solution set by our customer. Moreover, there are three discrete phases on the timeline that dictate our project planning more than the deliverables.

The project is divided into four overlapping work packages:

- **Project planning and background research** starts from the beginning
- **Project initiation** was launched when we received initial information about the available data
- **Technical execution** was launched when we received the actual data
- **Project management** is a work package that overlaps all of the work packages

After the project plan report, the initial schedule has been adjusted. Main structure is the same, but there was an unexpected delay (possibly related to the COVID-19 crisis impact on the client side) in compiling and encrypting the data on the client end. Furthermore, the first batch of data was corrupted. Eventually, approximately 30 Gigabytes of data was sent over a sequence of 60 encrypted 2-step verified emails.

3 Updated Risk Management Plan

Table 3 shows the updated table over the risks of the project. At this point of the project the team has received and familiarized themselves with the client data. Since the quality of the data has been of fairly bad quality, the risk of poor data quality has been upgraded to "High". Further, The team members have worked together now for an extended period of time, and they all have shown good communication skills and they are committed finalizing the project with the best outcome for the client. Thus, the risk level of insufficient communication between team members has been lowered to "Low".

Risk	**Probability**	**Effect**	**Impact**	**Mitigation Strategy**
Poor data quality	High	Misleading, incorrect or inaccurate results	High	Careful handling of data
Model too complex for the scope of the course	Medium	Too wide problem to solve for the allocated time	High	Focusing on explicit project goals.
Data security	Low	NDA contract violation	High	Local data management, risk assessment preceding deadlines
Insufficient communication between team members	Low	Resentment due to imbalance in workload between team members, misunderstandings	Medium	Regular communication between team members and manager and scheduling
Insufficient communication between team and client	Medium	Client not satisfied with the solution	High	Regular communication with the client
Team member inactivity or dropout	Low	High workload for other team members	High	Good communication between the project manager and the rest of the team. Clear schedule.
Resulting model does not provide accurate enough results	Medium	The tool will provide low or no value for the client	High	Strive for performance

4 Appendix

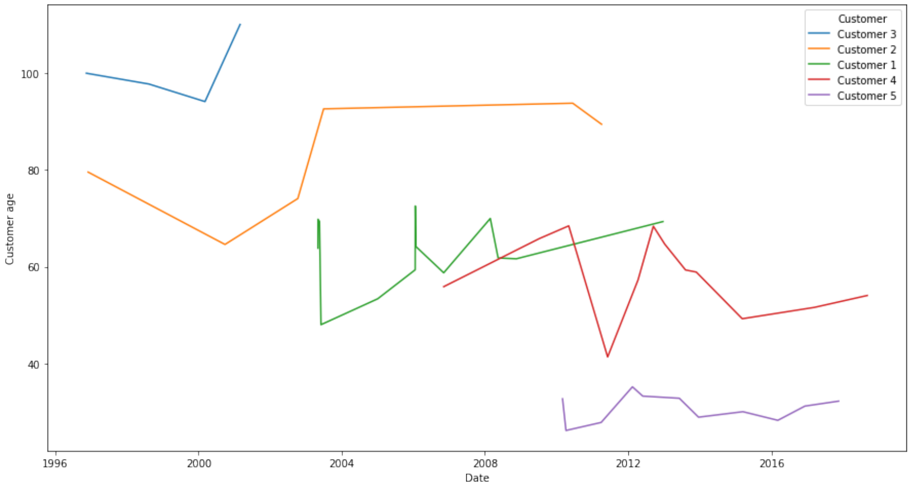

Figure 4: Age for 5 different random customers.

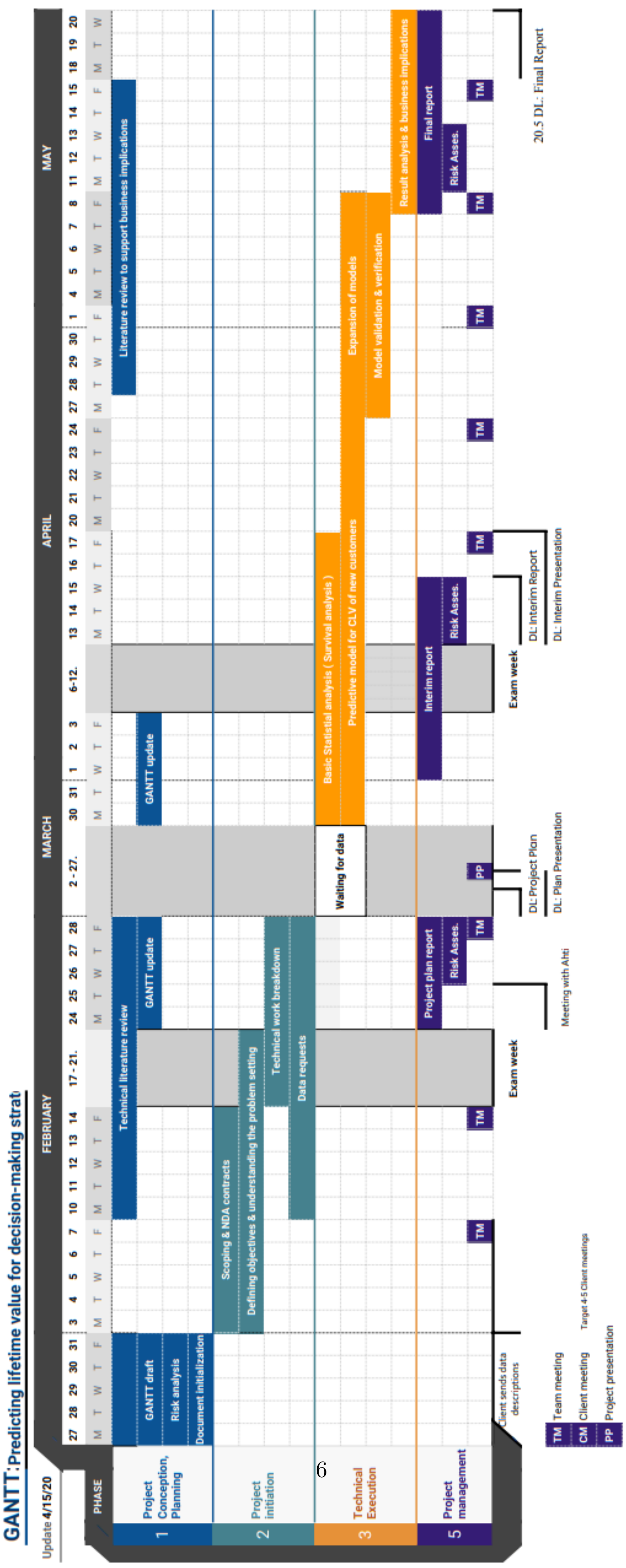

Figure 2: Project GANTT

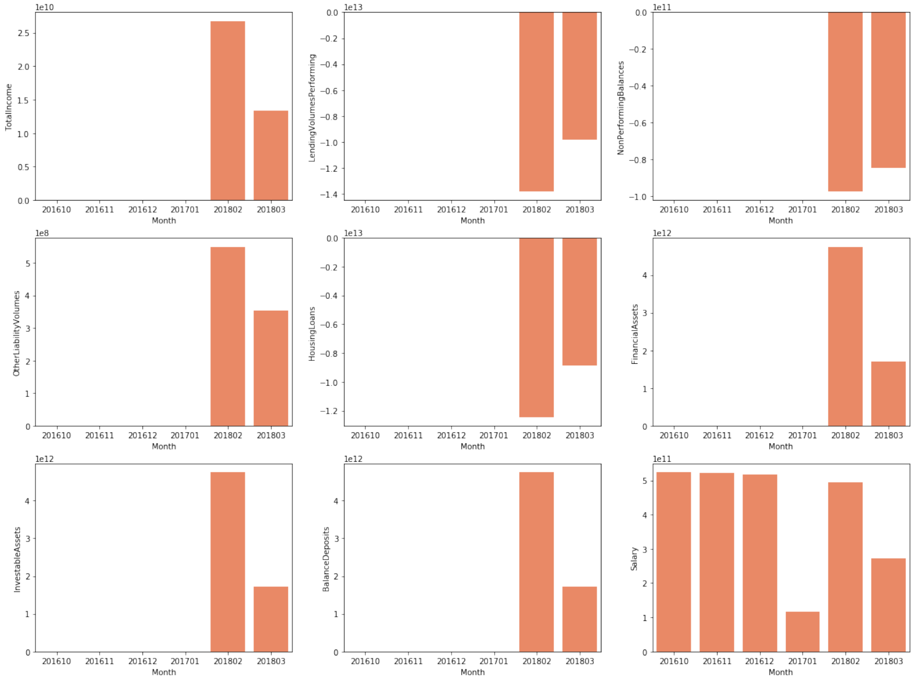

Figure 3: Total sum of attribute values for time series data attributes per month from a relevant subset.

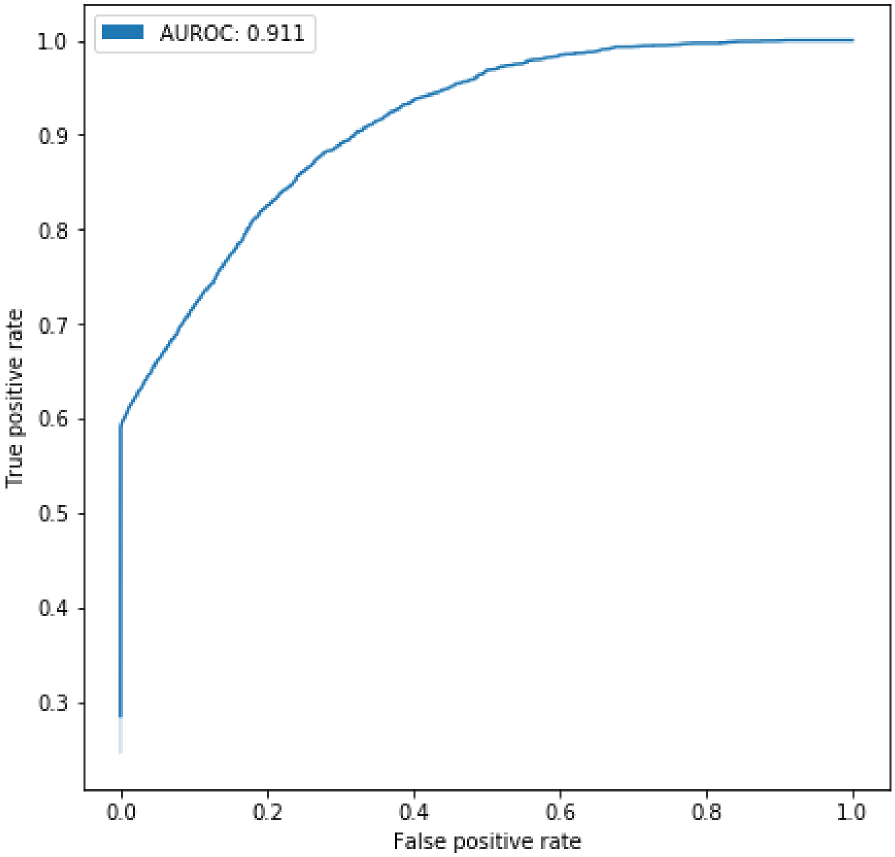

Figure 5: ROC curve of the classifier.

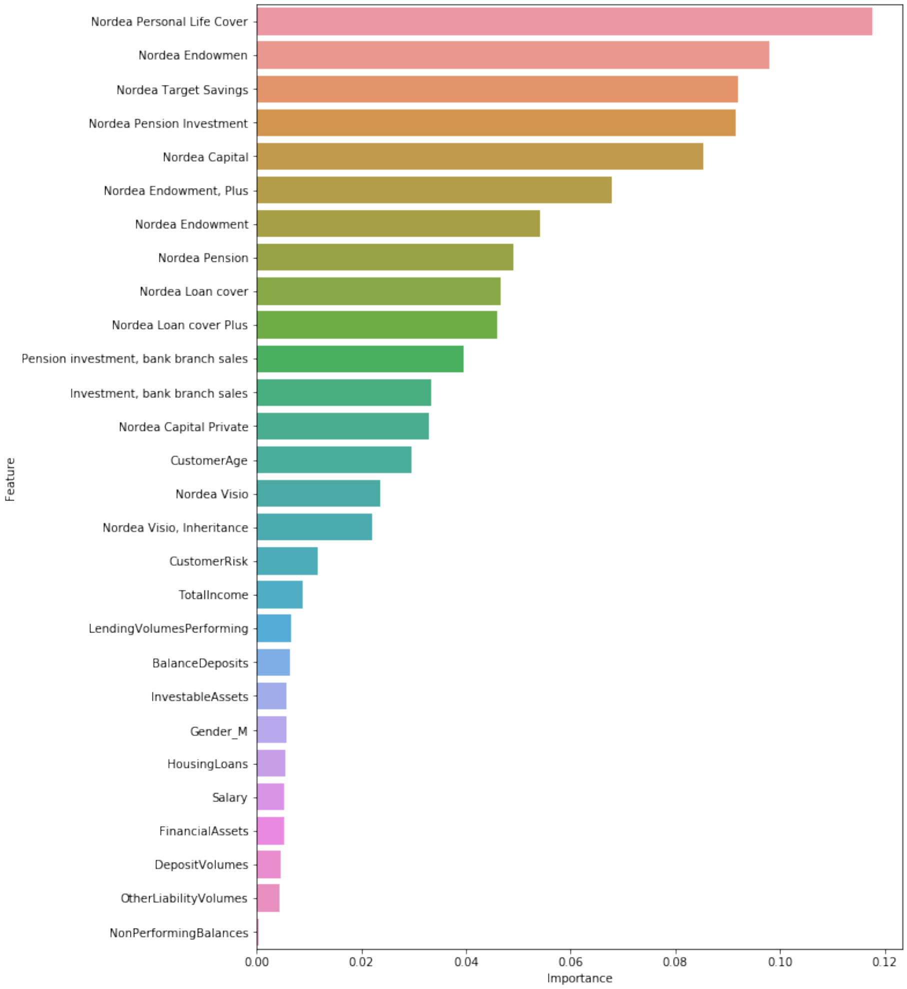

Figure 6: Extracted feature importances of the classifier.