Capstone Project Summary

Name: Miklós Márton Bóna
CEU Student ID: 1901545
Program: MS in Business Analytics
University: Central European University

This document aims at summarizing the Capstone Project I have conducted as part of my studies at MS in Business Analytics program at Central European University.

Due to the Non-Disclosure Agreement signed with my partner Company ("Client"), I omit all details from this document that could make the Client identifiable for the reader.

Introduction

My project was conducted at an industry partner company, with the aim of gathering insights from their data collected during their operations. The Client has expressed their interest in two particular target variables; fees per day (regression problem) and customer notification (binary classification problem).
The goal of this project was two-fold; first, to conduct a thorough exploratory data analysis (EDA) to clean and understand the data structure and second, to develop supervised machine learning (ML) algorithms to identify the most impactful variables that contribute to the target variables.
For the EDA as well as for the modelling, I used R programming language and worked in R Markdown files to ensure reproducibility.

Exploratory Data Analysis

The raw data is sourced from the Client's internal database and was transferred to me via cloud. The quality of the raw data is satisfactory, however multiple data transformation steps were required to make it fit for modelling purposes.
I was working with two separate databases; (1) one that contains all events throughout the observation period and (2) one that contains each notification associated to said events submitted to the Client.

Data transformations

As stated above, certain data transformations were necessary to fit the raw data for modelling. Before each transformation, I evaluated the specific business case, potential impact and other alternatives as well. In general, my aim was to only drop variables or observations from the data if I had no other option.

The list of data transformations is presented below:

1. **Too many missing values**: replace NAs or omit observations with missing values or omit variable
2. **Transforming variables to factor variables**: character-to-factor or numeric-to-factor
3. **Not enough variation in explanatory variable**: group variable factor together with another factor or omit variable
4. **Manual override**

After data exploration and data cleaning, I joined the two databases. Target 1 prediction only uses explanatory variables from the first – event level – database, while Target 2 uses both databases. The join required to apply a certain de-deduplication logic on the second database, so that the target variable remained unique. To uncover additional information, I applied feature engineering on the raw variables present in the joined data.

Overall, there are more than half-a-million observations in the joined and cleaned dataset

Modelling

Methodology

To train supervised ML algorithms, I used a stepwise model building approach: first, I categorized the potential explanatory variables into groups, then I started building models with increasing complexity – by adding the above variable groups to the list of explanatory variables – so that I can monitor each group's contribution to the predictive power.

I have used two types of supervised learning techniques; Generalized Linear Model (GLM) and Elastic Net Generalized Linear Model (GLMNET).

Multicollinearity between the explanatory variables is expected to be a problem with GLM models, as it does not use any regularization techniques, as opposed to GLMNET. To tackle this, I used a stepwise method to reduce Variance Inflation Factor (VIF) among the explanatory variables. The approach was to test VIF on a trained model, then check each variables' VIF-score, and decide whether or not to omit that variable. Multicollinearity may not be a significant problem, if a VIF-score of a variable is below 10. GLMNET combines the advantages of lasso and ridge regressions which help dealing with multicollinearity, by reducing the coefficient weights of collinear explanatory variables.

Target 2 is a binary classification problem, where the data is highly imbalanced; about 6% of the observations belong to Category 1, and the majority of the cases to Category 2. In order to treat this imbalance, I experimented with *upsampling*; a method that creates artificial observations in the minority category, so that the classes would be balanced in the train sample.

I sampled 10% of the data in order to speed up model training, and then applied the best model – for each target – on the full dataset.

Results and conclusion

Overall, the analyses conducted shed some light on the most impactful variables for both targets. The explanatory variables with the highest absolute coefficients can be logically interpreted form a business point of view, therefore the analyses may find support for further research from the Client. On the other hand, I was not able to increase overall model performance for neither of the two targets.

My project did not cover these, but more complex ML algorithms (e.g. tree-based methods) may produce more accurate predictions. Another suggestion for further research would be to fine tune certain data transformation steps (mainly the re-grouping of class variables) I took, in order to avoid overfitting to the train data. This may reduce the probability of overfitting, but important information from the raw variables could be lost.

My main focus with this project was to conduct the research as reproducible as possible, as well as to provide a very thorough technical specification document – along with the R scripts – for the Client, in case they decide to continue the project.