Assignment 2: Predictive modeling

Group 11: Roger Xavier Granda Quichimbo, Maria Andrea Martens,
Alexandre Pitsaer, Knarik Sahakyan

Advanced Analytics in Business, KU Leuven
Prof. Seppe vanden Broucke
{rogerxavier.grandaquichimbo,mariaandrea.martens,alexandre.pitsaer,
knarik.sahakyan}@student.kuleuven.be

1 Remarks on the code provided

Together with this report we provide two Jupyter notebooks:

- **Data PreProcessing and Exploration.ipynb** : This notebook contains the code for the data prep-processing, exploration, and visualization.
- **Model Testing and HyperParameter Tunning.ipynb** : This notebook shows the training phase of our pipeline. It also includes the code for hyper-parameter tuning of several models and our feature selection strategies.

2 Introduction

Machine learning has experienced tremendous progress in the last two decades and its use has become ubiquitous in marketing and customer management.[1]. Customer churn prediction is one of its application that helps firms to sustain their profits, developing more efficient and more effective strategies to retain their most valuable customers. [2–4].

According to [5], acquiring a new customer is 5 to 25 times more expensive than retaining an existing one. This is a critical issue in telecommunications, which is a highly competitive sector. The purpose of this project is to develop the whole pipeline that is required to develop machine learning models that can predict churn of wireless telephony customers of a telecom firm in South America.

This report is organized as follows. First, we describe the data pre-processing and data exploration phase. Second, we present the model training phase and the challenges that we encounter. Finally, we conclude and analyze what we learnt from this project.

3 Data Pre-processing, Exploration, Transformation

An initial training dataset containing 5K records was used. This original dataset contains 38 features. Examples of features are: ”churn label”, if a client has a data plan or not , ”start date of subscription”, ”last month data consumption”, ”last month sms sent”, etc. Figure 1 depicts a summary churning/not churning clients that *have/don't have* a data plan with the telecom company. During the initial data exploration phase, different plots (box plot, cat plots, histograms) were generated to visualize and understand the data distribution. *Start date of subscription* was transformed into a numerical value (*duration* feature) which

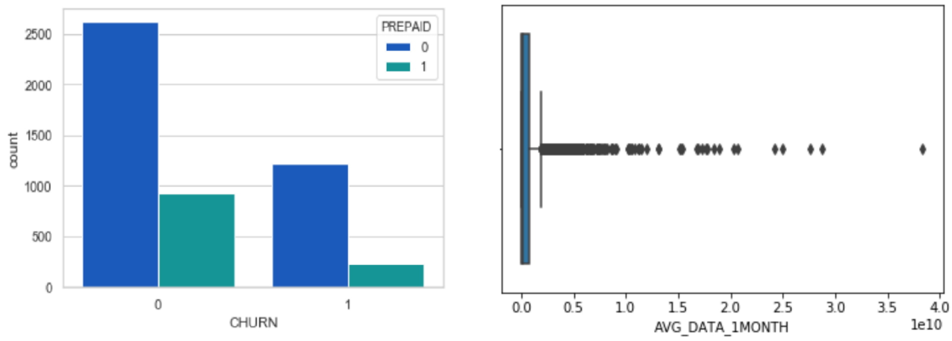

Fig. 1. (left)Number of churning/not churning customers aggregated by type of subscription in training data. (right) Sample of outliers in average data usage feature

measures the number of days that the customer has had a subscription with the company. Furthermore, data plots revealed outliers. Figure 1 shows a sample of outliers identified in one of the features of the training set. We decided to drop these outliers before training the model. Furthermore, three variables of data consumption showed null values; to solve this issue, a zero value was imputed and a new variable that described whether a customer had a data plan or not was created. In addition, we decided to create two new features to enhance the model's performance in later stages. *Call Network Size in Weekends* and *SMS Network Size* were created to differentiate between business and private customers.

We chose to partition the data for the training and test set in the following proportions: 80% for training, 20% for testing. After using plots, feature re-scaling was applied; in particular, we Robust Scaler to modify the scales of each numerical feature. PCA analysis was performed to find representative features using re-scaled features. We found that two principal components explained more than 85% of the total variance of the data. These principal components were added as new features. To further explore the data, we used TSE feature Visualization, figure 3 depicts churn data using all features except ID and dropped features.

The resulting subset of features were considered for the model's training section:

prepaid, fin_state, count_payment_delays_1year, days_payment_delays_current, days_payment_delays_1year, complaint_3months, count_offnet_calls_1week, count_onnet_calls_1week, avg_data_3month, count_connections_3month, avg_data_1month, count_sms_inc_onnet_6month, count_sms_out_offnet_6month, count_sms_inc_offnet_1month, count_sms_inc_offnet_wkd_1month, count_sms_inc_onnet_1month, count_sms_out_offnet_wkd_1month, count_sms_out_onnet_1month, avg_minutes_inc_offnet_1month, avg_minutes_inc_onnet_1month, minutes_inc_offnet_wkd_1month,minutes_inc_onnet_wkd_1month, avg_minutes_out_offnet_1month, avg_minutes_out_onnet_1month, minutes_out_offnet_wkd_1month, minutes_out_onnet_wkd_1month, plan_duration_days, year_suscription, has_data_plan, pca1,pc2

At the end of this data exploration phase, the following assumptions were stated for the next steps in the project:

– The test will only be performed on December 2013. External variables such as macro-economic indicators will not be taken into account.

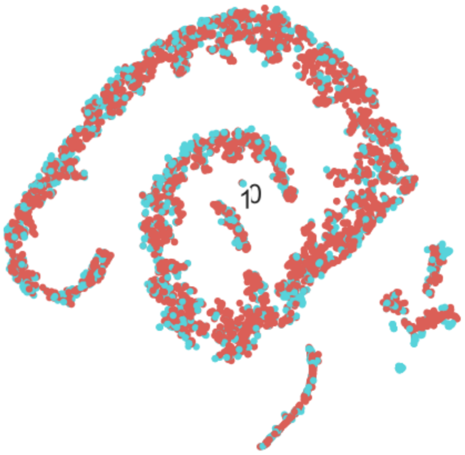

Fig. 2. TSNE feature visualization. 1: Client churn 0: Client remained at company

- The complaint data for one week and for two weeks seems to be faulty. When comparing the number of complaints for one week, two weeks, and one month, the aggregates do not match in around 200 cases (where there are more complaints in the last two weeks than during the last month) and around 100 cases (where there are more complaints in the last week than in two weeks). Hence, we will focus on one month, three month, and six month data for the complaints.
- Given that the financial state variable is almost perfectly correlated with the churn variable (except for 69 observations in the whole dataset) and be believe that it might just be set administratively to a value when a customer churns, meaning that it would not be predictive per se, we will exclude it initially.
- We will initially drop customer life value given that it has outliers and it is not clear how this variable was computed.
- We will infer a new binary feature, which will inform whether customers have a data plan or not.

4 Model Training

The following models from the Scikit Learn models were used:

- ○ Logistic Regression
- ○ Random Forest
- ○ XGBoost
- ○ Neural Networks
- ○ Voting Classifier (Ensemble Model): Random Forest+Gradient Boosting+Logistic Regression

We reported the performance of our model using AUC, i.e. the area under the ROC curve (false positive rate vs true positive rate). AUC represents the probability that our model gives a higher score to a randomly chosen churner than to a randomly chosen non-churner. Hence, this metric should get as close as possible to one. We do not use accuracy as this performance measure could be misleading in our case, since it assumes that misclassification costs are symmetric and it lacks a reference point when classes are not balanced. To test the aforementioned models, the initial set of features was used in the training phase. At this stage, the models did not perform better than 0.80 AUC for XGBoost. Next, class imbalance problems were identified (Churn: 2835, Not Churn: 1149 for training). To tackle this problem, Smote sampling strategy was used. Churn points were augmented to achieve class balance (Churn: 2835, Not Churn: 2835 for training). Other approaches were also used (oversampling, undersampling), however, performance decreased or did not perform better using those approaches. For hyper-parameter tuning a RandomSerch CV strategy was used to reduce the space range of search of hyperparameters. Subsequently, Grid-Search CV was used to find the best values for hyperparameters. XGboost obtained the highest AUC score; the best parameter values were: (n_estimators=700, learning_rate=0.1,subsample=0.8) To reduce the complexity of the models, Boruta Feature Selection was implemented. The resulting list of important features was described in the former section.

Finally, both models were tested using the list of features described in previous section and also the same subset + FIN_STATE feature. Initially, we decided to drop the Financial State feature because we were suspicious that it was automatically set when churn happened, therefore not having any true predictive value. Nevertheless, the AUC metric did not improve despite of several attempts of tuning the hyperparameters using cross validation. For this reason we decided to test our models also including the Financial State variable. Table 1 shows the results generated by models in the training phase. Furthermore, figure 3 depicts ROC curve for the model that used FIN_STATE feature for each class (Churner and Not Churner). One can notice that despite of the asymmetric nature of data set, both classes were predicted with successful performance. Given our suspicions about the Financial State variable, we submitted our model without it to the leaderboard. As already stated we do present results with and without this variable in the present report.

Table 1. Results after training several model for churn prediction

Model	**Initial Features Set (AUC)**	**Initial Features Set + Financial State (AUC)**
Logistic Regression	0.70	0.978
Random Forest	0.84	0.991
XGBoost	0.87	0.991
Neural Network	0.84	0.98
Voting Classifier	0.86	0.99

5 Conclusions

Churn prediction benefits greatly from the use of machine learning methods. We implemented several models that had an AUC performance of 0.98 or more with our optimal hyperparameter choice. Including financial state, a feature that we

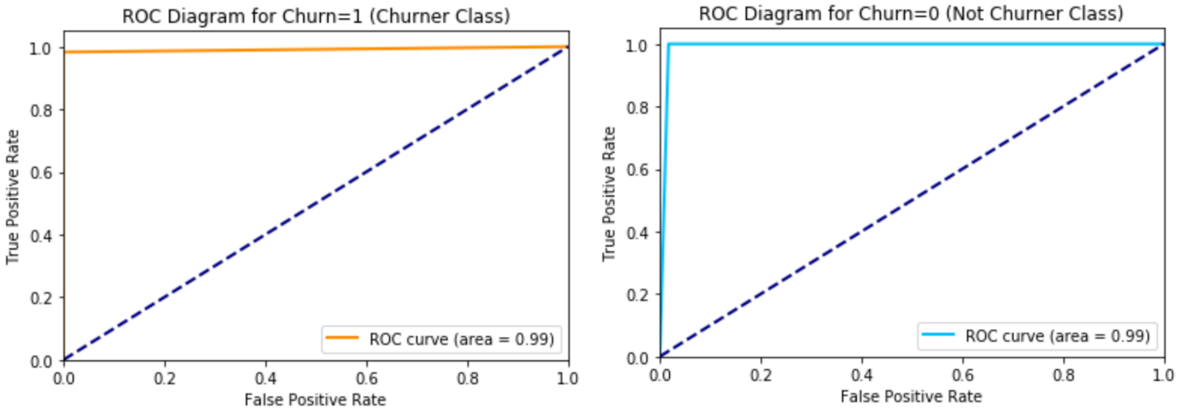

Fig. 3. ROC Diagrams of the best model found for binary predition. (1: Client churn) (0: Client remained at company)

did not fully understand and suspected to be very highly correlated with the variable we wanted to predict, noticeably boosted the performance of models. We were skeptic on reporting the scores on the public leaderboard including this financial state for this reason. Our report does report the scores of our models including/excluding financial state.

References

1. M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.
2. A. Keramati, H. Ghaneei, and S. M. Mirmohammadi, "Developing a prediction model for customer churn from electronic banking services using data mining," *Financial Innovation*, vol. 2, no. 1, p. 10, 2016.
3. T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," *Simulation Modelling Practice and Theory*, vol. 55, pp. 1–9, 2015.
4. A. Amin, B. Shah, A. M. Khattak, T. Baker, S. Anwar *et al.*, "Just-in-time customer churn prediction: With and without data transformation," in *2018 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2018, pp. 1–6.
5. "The value of keeping the right customers," https://hbr.org/2014/10/the-value-of-keeping-the-right-customers, accessed: 2019-05-28.