Machine Learning Engineer Nanodegree

Capstone Proposal

Miloš Ranđić, Telenor Serbia

September 5, 2017

Domain Background

Nowdays, telecommunications industry is the phase where customer needs are in the center of attention. The greatest challenge is to align specific product offers to precise group of customers, in order to make customers satisfied with existing products and services, and to maximize company's revenue as well. Since the focus is on the customer, CRM(*Customer Relationship Management*) is one of the main points of interest for telco industry, especially when applying with machine learning[1].

It seems that over the years customer data has been collected and increased exponentialy in its volume, which means the user behaviour can be tracked and can teach us how customers think, what their needs are, and based on those rules, encourage us to conclude what product they should be offered. Telco industry is just a great place to make a research of user behaviour and create various product offers using supervised machine learing techniques. Supervised learning as a concept has been in the industry for a while, but nowdays its applications are growing exponentially[2].

Personal motivation lies in the need to make a research on behaviour of customers in telco industry, in order to build machine learning model for product recommendation with aim to generate more profits from customer services consumption[3].

Problem Statement

For this project, I have chosen the challenging problem about offering roaming add-on services to customers who travel to foreign countries. Since the roaming traffic is heavily billed while using international telco operator services, Telenor ensures that users who activate roaming add-on service won't be billed for additional fees as long as they consume mobile traffic within the add-on package. Of course, customers pay additional fee for this product at the moment of activation of the product, but the price is far cheaper than using mobile traffic without the add-on. Database that will be proposed tracks two groups of customers, on a monthly basis: users that buy this product when travel abroad and users that don't buy this product when travel abroad.

The goal of this project is to build machine learning model that could solve the problem of defining which users are more likely to buy this product in order to send them an offer for activating the service. On high level, the potential solution would consider building machine learning model using available data with appropriate tests using test data, in order to examine the performance of the trained model.

Since this is a two-category classification problem, one of the evaluation metrics that will be considered is confusion matrix, from which will be derived different metrics for measuring performance (precision, recall, accuracy and F-score).

Firstly, naive predictor should be built, using descriptive statistics in order to build naive classification critetia.

Datasets and Inputs

The dataset considers *Telenor business customers* behaviour data available on a monthly basis. Data is available for a period of 16 months (04/16 to 07/17) . **Important notice:** Identity of the users and their true behaviour are **anonymized**, for privacy and user protection purposes declared by Telenor. Entire usage/revenue data presented in this dataset has been scaled using unique secret coefficient, avoiding privacy policy to be affected in that way.

Dataset contains following variables:

VARIABLE NAME	DESCRIPTION
TIME_ID	Year-Month of recorded behaviour.
SUBSCRIPTION_ID	Unique ID of a user.
SERVICE_SEG	Segmentation of user. There are four ordinal segments: N(New), Br(Bronze), Si(Silver), Go(Gold).
AP_NAME	Type of tariff plan. There are two possible roaming tariff plans: Travel Sure 10 and Travel Sure 50.
ACT_MONTH	Month of activity. This variable indicates the month [1..12] of the user being in roaming.
ACT_VOL	Entire monthly usage generated within international network, measured in points (mixed SMS+VOICE+INTERNET).
REV	Entire monthly revenue generated within international network.
NUM_OF_MONTHS	How many months in the past 4 months the user has been in roaming(within international network). Possible values: [0..4]
VOICE_VPN_MIN	Total monthly VPN minutes spent.
VOICE_ONNET_MIN	Total monthly minutes spent inside Telenor network.
VOICE_OFFNET_MIN	Total monthly minutes spent outside of Telenor network, towards national networks.
VOICE_INT_MIN	Total monthly minutes spent towards international networks, within national network.
SMS_NAT	Total number of SMS messages sent towards national networks, within national network.
SMS_INT	Total number of SMS messages sent towards international networks, within national network.
GPRS_NAT_MB	Total number of MBs spent within national network.
OS	Operating system user runs on a mobile device.
MF_REV	Total fixed monthly revenue from package bundle.
US_REV	Total monthly revenue from over the bundle usage.
FLAG	Is the user taker or no_taker of the add-on in particular month. Possible values [TARGET, NO_TARGET]. This is our target variable.

For each month, data is provided using described features. For categorical features, one-hot encoding preprocessing will be done. Missing values are not present in this dataset. Additional pre-processing techniques will be conducted, including feature transformation(log-transform for skewed distributions) and normalization (scaling), in order to prepare data for learning algorithms. Considering the amount of data, 12 months of data will be used for training of the model and the rest of the data will be used for testing purposes. The dataset was obtained by extraction from data warehouse using SQL query language from BI department in Telenor Serbia. Selected features in the dataset relate to this problem because those features cover different spans of usage and behaviour, both within national network and in our case more interesting, international networks.

Solution Statement

Solution to the problem lies in following

Preprocess dataset – perform all neccesary preprocessing and feature engineering. This includes feature transformation and scaling of our data. Dataset should be split into train and test set.

Model design – Initially, naive predictor should be built, using classification criteria obtained from visual observations of dataset. Accuracy and F-score should be calculated from naive predictor on test set. These scores will be compared to scores calculated from learning algorithms. Considering learning algorithms, appropriate model among different types of classification models should be chosen. From selected models, the one that performs best on the test set should be used for further model improvements.

The model with highest performance on the test set should be used for generating campaigns for targeting users to buy add-on service. This campaign test will be used for demontration purposes only, since this model is not in the production yet, hence, it does not affect user behaviour from which we could measure our true campaign success rate. Only as-is success rate, without impact of our model will be observed. Take rate and roaming rate (explained in following sections) will be evaluation metrics for our campaign success.

Benchmark Model

Firstly, the most effective model among three learning algorthms will be compared against naive predictor. The most effective model should be further optimized in order to maximize its performance on the test set. At the end, three groups of scores should be compared: naive predictor score, unoptimized model score and optimized model score.

Omptimized model will be used for generating campaigns for offering add-on to users. The idea is to offer add-on to users in month *t+1*, based on users that learning algorithm selected as buyers in month *t*.Take rate will be measured from the target list. Roaming rate should be considered also, regarding the percentage of targeted users in month *t* that actually have been in roaming in month *t+1.* There will be generated 3 campaigns in total. Take rate and roaming rate will be observed.

Evaluation Metrics

For this project, Accuracy and F-score measures will be used as evaluation metrics. Using these metrics naive predictor solution will be compared to learning algorithm soultion. Precission and recall will also be considered as evauation metics in order to track how our algorithm correcltly classifies users to one of two classes in test set.

$$precision = \frac{TP}{TP+FP};\ recall = \frac{TP}{TP+FN};\ accuracy = \frac{TP+TN}{TP+TN+FP+FN};\ F_\beta = (1+\beta^2)\frac{precision*recall}{(\beta^2*precision)+recall}$$

Selected optimal model will be used for creating target list in month *t*, in order to predict the potential buyers for the month *t+1*. Users who are selected by model as buyers in a month *t* will be assumed to buy the product in the next month, *t+1*. Take rate is calculated as ratio of actual takers intersected with target list, and actual roamers in month *t+1*. In this way, we calculate take rate as a ratio of users who activated add-on from target list and users from target list who were actually in roaming in the next month. Roaming rate explains the percent of correctly predicted users from target list, that happend to be in a roaming in the next month.

$$takeRate = \frac{count(targetedUsers_t \cap takers_{t+1})}{count(targetedUsers_t \cap (takers_{t+1}+noTakers_{t+1}))};\ roamingRate = \frac{count(targetedUsers_t)}{count(takers_{t+1}+noTakers_{t+1})}$$

Project Design

This project will have following workflow:

Dataset exploration

This section is about getting familliar with the data. Overall statistics from dataset will be shown, including total number of records per each class, etc. For variables from dataset, relative frequency histogram charts will be shown, comparing the differences between two classes we want to model (takers and no_takers), in order to visually track differences between these two classes.

Building naive predictor

The aim of previously described comparison is to build naive predictor using criteria obtained from visual observations of dataset. Finally, naive predictor will be tested on test set form date range 04/17-07/17. F-score and accuracy score will serve as a comparison metrics to scores calculated from learning algorithms.

Data pre-processing

In order for learning algorithms to perform properly, data pre-processing needs to be done, for both continous and categorical features. For categorical features, One-Hot encoding will be done, in order to transform categorical features into numerical features. For numerical features, each one will be checked for skewed distribution. If so, the feature will be transformed using the log-transform function. At the end, all continous features will be scaled to [-1,1] range.

Train machine learning algorithms

Three learning algoritms will be chosen for this project: Gaussian Naive Bayes, Random Forest and Gradient Boosting algorithm. These three algorithms will be trained using default parameters on the same train set from data in range (04/2016-03/2017). Since the dataset is not balanced, rebalancing will be done by downscaling the larger class to a lower percentage number of records. There will be three scenarios: two, that use downscaled, balanced dataset and one that use original balance ratio, without downscaling being performed. Model training time will be tracked for each scenario. Models will be tested on small portion of training data and on a test set from data in range (04/2017-07/2017). For both training and testing data, F-Score and accuracy measures will be provided. The model that performs the best on the testing data will be chosen. The chosen model preformance metrics on the test set should be compared to performance metrics of naive predictor.

Optimizing the learning algorithm

Now when we have our model, the optimization process should be performed. Model should be trained using different training parameters with possible downsampling of train set. Feature importance defined by optimal model should be listed.

Final benchmark

In this section, there will be three pairs of F-Score and accuracy metrics: for naive predictor, unoptimized model and optimized one. The expectation of this benchmark is to outline that the optimized model should perform best comparing to naive predictor and unoptimized model performance.

Demontration on making campaign offers

This additional section should provide an insight on how this trained model could potentially be used in practice. Since we did not apply this algorithm for making real offers, we cannot track real effect of the campaign success. In that way, with this algorithm we can only track as-is situation, and measure the effect without machine learning algorithm being applied. In that way, we can measure as-is take-rate scores and expect that, when the model prediction from this project is applied, the take-rate should be higher.

In order to make target list for the month *t+1*, users that were selected as buyers by machine learning algorithm in month *t* should be selected as potential target for month *t+1* (these are TP and FP groups of users from previous month, if we look at confusion matrix). Take rates from TP and FP groups should be observed separately, in order to see in what percent users that activated the add-on in month *t* (TP group) activate add-on in the next month as well. Users from FP group should be considered as potential targets (new buyers, since this group of users did not activate the service in previous month). FP group of users could be a measure of campaign success, if we maximize number of takers during real campaigns using machine learning. One additional rate that will be considered is roaming rate - percentage of users from target list from month *t* that were in roaming in month *t+1* (not minding if that user has activated service or not). Since we have 4 months of test data, 3 campaigns will be created (1-2, 2-3, and 3-4 month pairs) in order to test how our model predicts potential buyers for the next month. Model prediction scores of target lists should be plotted accordingly.

References

[1] M. Cioca et al., *"Machine Learning and Creative Methods Used to Classify Customers in a CRM Systems"*, Applied Mechanics and Materials, Vol. 371, pp. 769-773, 2013

[2] Hastie T., Tibshirani R., Friedman J., *"Overview of Supervised Learning",* The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York, NY, 2009

[3] L. Zhao et al., *"An Overview of the Recommender System"*, Applied Mechanics and Materials, Vol. 302, pp. 787-791, 2013