Credit Card Usage Prediction

Tamas Veress, 2016 Sep

Abstract. This paper presents various methods to predict bank card users' behavior for the purpose of credit card upselling. We extract features from both user data and activity time series and also derive features capturing the popularity of branches. We apply eXtreme Gradient Boosting (XGBoost [1]) binary logistic regression and blend different approaches to improve our prediction.

1 Introduction

The Bank Card Usage Challenge of the European Conference on Machine Learning [2] asked the participants to predict the behavior of the Hungarian OTP Bank's clients. The task was split into two parts. Firstly, we had to predict which branches the clients will visit during the second half of 2015. Secondly, we had to forecast which one will apply for credit card. This paper presents the approaches that achieved the third place on the public leaderboard of the first task and second place on the latter one.

The paper is organized as follows. The next section provides the data processing steps and models of the branch visit prediction task. The third part describes the models used to forecast the credit card application. Finally, the results and brief discussion presented.

2 Bank Branch Visit Prediction

2.1 Data Preprocessing

We can interpret the evaluation metric of this task in the way that we are not really interested in how many times exactly the users go to each branches but we rather curious about the preferred branch of each user. Thus one of the goals of the data preparation is that we convert the raw transactions to a target and feature set of a binary logistic regression problem.

Another goal of the preprocessing is to reduce the huge user and branch space to a smaller one to decrease the run time of the subsequent optimization algorithms.

Both our training and test data set contains transactions of above 190 thousand users and the training data includes branch visit data to 323 branches, too. We expect users to visit branches closer to their homes or other places they regularly visit. While ranking the distance between the branch and user location is straightforward we need further work to express the distance between multiple shopping locations and the branches. We

choose the simplest way by taking the median of the geographical coordinates of the shopping points for each clients and calculated the distances between the median shopping location and the branches. After calculating the ranks, we kept only those user and branch pairs for which at least one of the following conditions holds:

1. The branch is at least the 5th closest to the user's home
2. The branch is at least the 5th closest to the user's median shopping location
3. The branch was visited by the user at least once in 2014

These restrictions resulted a relatively small training and test data set with 1,220,225 and 1,189,072 records respectively. The appendix presents the list of variables generated by the preprocessing script along with their description.

We introduced derived features with Hcross and Pcross to capture the varying marginal impact of distance and rank. As Numar et al. [3] have found analyzing restaurant choices people tend to choose the closest restaurant with lower chance if they live in a dense city with abundant restaurants nearby.

2.2 Model

We applied XGBoost due to its speed and accuracy on categorical feature set. Table 1 shows the feature importance of our first level model.

Table 1. Feature importance of the first level model.

Feature type	Feature	Gain
Distance features	PDISTrank	0.3154
	HDISTrank	0.2430
	Hcross	0.1186
	Pcross	0.0949
	HDIST	0.0932
	PDIST	0.0657
User features	CIT	0.0271
	CAP	0.0256
	VIL	0.0117
	Age1	0.0031
	GEN	0.0011

We can see that distance related features are the most important drivers in branch choice. User attributes has little role but we kept them as both local cross validation and the leaderboard supported their robustness. We expect that CIT, CAP and VIL still capture some of the above mentioned relationship between branch density and choice. We interpret the impact of Age1 and GEN with mobility. Younger males tend to be more mobile decreasing the chance to choose the closest branch.

Intuition suggests that beyond distance and user attribute branch choices are impacted by other factors too. Some branches might be more conveniently located at transport hubs or they offer different or broader range of services. In order to incorporate this into our prediction we derive additional features that quantify local and global popularity of the branches.

EXCESS, our global popularity feature was derived by taking the average error term of the first level XGBoost model for each branches. FREQ, TOP and RANK are the local popularity features. We generated these by calculating how many users from the same location visited the branch and ranked those frequencies for each location. We rounded the geological coordinates to kilometer to create location identifiers.

We included these new features into the final model along with the predicted values of the first XGBoost model. To control information leakage we split the training data into two parts and generated PRED and EXCESS for each by training the model on the other one.

Table 2. Feature importance of the blender model.

Feature	Gain
PRED	0.6821
FREQ	0.1986
EXCESS	0.0504
TOP	0.0383
RANK	0.0196
CIT	0.0045
CAP	0.0038
VIL	0.0015
GEN	0.0008

Table 2 presents the feature importance of our final model. Local popularity features seem to be strong factor but global popularity has some explanatory power too. The stacked model's performance on the public leaderboard was 0.6744 while the simple XGBoost achieved 0.6447.

3 Upselling Prediction

3.1 Data

Credit cards offer wide range of benefits from which one of the most important is the percentage cash-back on all or certain purchases. That is why we expect that credit card application and credit card ownership is correlated with income, wealth, the amount and frequency of card usage. The appendix shows the list of features we generated for our models.

We tried to capture the benefits of the credit card with some more focused features too. OTP offers extra cash-back at partner fuel stations. We suspect that fuel stations are denoted by market type 'b' thus we calculated the amount spent for only this market category (AMTb). We thought POSSD could serve as a good proxy for this too. The more the user travels across the country the more likely she burns more fuel thus would be worthwhile to have a credit card.

Beyond the cash rewards credit cards often include travel insurance that can be highly valued by those often traveling abroad. We made an attempt to capture this with the number of transactions made near to the Liszt Ferenc Airport.

We also expect different likelihood of application depending on credit card application or ownership history.

For most of the features we produced both the six month and yearly versions so that we can test our models with making both user and time split in the training data.

3.1 Models

In our final submission that earned the second place on the public leaderboard we combined three different model predictions. In the three models we applied XGBoost, however, the feature set, target variable and training set was slightly different.

In Model 1 our target variable was the credit card application during the second half of 2014, features were related to the first half year transaction. We used the Jan-15 – Jun-15 application data and Jul-14 – Dec-14 features of the same users for cross validation. Table 3 presents the feature importance of the model.

Table 3. Feature importance of Model 1.

Feature	Gain
AMT	0.2200
INC	0.1792
PosFreq	0.1731
NetFreq	0.1251
AMTb	0.1035
Have	0.0563
Age1	0.0225
GEN	0.0193
WH	0.0179
CIT	0.0175
Age2	0.0169
VIL	0.0152
CAP	0.0148
Age3	0.0123
Had	0.0057

We can see that income and transaction features are the most important determinants of credit card application. User location, age and gender are considerably weaker.

Model 2 was developed to address the imbalance in the dataset by using Have as target variable. The idea behind is that looking at which users actually have credit cards and predicting who 'supposed to have' credit card can be useful for predicting who will actually apply. We expected that this approach will reduce the problem coming from the imbalance since the number of credit card holders are around 7 times higher than the new applicants in a year. On the other hand, the relationship between the features and target can be different in the two models. In our training data we do see that income is more correlated with the application than with the holding. We hypothesize that this

is because when someone apply for credit card they have to have income above a certain level while there is no income threshold for keeping the credit card.

Overall we found that Model 2 improved marginally both our local CV and the leaderboard score if we included with 0-10% weight thus we used it for more submissions. During the blending we only used Model 2 for users who do not have credit card.

Finally, we introduced Model 3 which is a version of Model 1 however we used user split instead of time split for cross validation and we also expanded the card holding features. The motivation behind was that we expected that we can find some strong relationship between application and card holding history. Those who have credit card for long time are expected to apply with more chance compared to those just obtained one simply because there is higher chance that their card would expire in the subsequent six months.

On the other hand, Model 3 turned out to be very sensitive to the model parameters due to the high imbalance and low sample size. We also suspected that our model gives biased result because in the training set only the first application date is included. Nevertheless, we included it with 1% weight in our final submission reaching further improvement in our public leaderboard score.

4 Conclusion

We saw that from the perspective of branch choice distance related features are the most important drivers. We found strong evidence that customers tend to prefer closer branches and rank has stronger explanatory power compared to the pure distance. The accuracy of prediction improved by including some user specific features: age, gender. Beyond this we were able to generate features that capture the popularity of certain branches. Local popularity turns out to be more prevalent compared to global popularity.

Potential improvement areas lie behind our restrictions. Probably the strongest restriction is that we excluded those branch user pairs that are relatively far from each other. Around 20% of the branch visits are made beyond the 5th closest branches that is why we expect that our result would improve significantly with relaxing this restriction even if we consider that we partly solved this issue by introducing local popularity features. It would be interesting to see how much real improvement we achieve with blending after running the first level XGBoost without this limitation.

The median shopping location is another important simplification. This can give very misleading result when the user's shopping locations are concentrated in more cities which are not close to each other.

In the upselling prediction we found that income and various historical card transaction features are strong factors affecting credit card application. We attempted to tackle the issue from imbalance in our data set by developing multiple models and combining those predictions.

We believe that all of our three models have valid contribution, however, we would expect improvement by building more systematic local cross validation to validate the stacking of different models.

Appendix: List of Features

Feature	Description
Branch Visit Prediction	
TARGET	1 if user visited the branch during 2014, 0 otherwise
HDIST	Distance between user address and branch
HDISTrank	Rank of HDIST with 1 showing the closest branch
PDIST	Distance between median shopping location and branch
PDISTrank	Rank of PDIST with 1 showing the closest branch
Hcross	HDIST* HDISTrank
Pcross	PDIST* PDISTrank
CIT	1 if user located in a city, 0 otherwise
CAP	1 if user located in the capital, 0 otherwise
VIL	1 if user located in a village, 0 otherwise
Age1	1 if age is under 35, 0 otherwise
GEN	1 if male, 0 otherwise
Upselling Prediction	
Variables	Description
AMT	Amount spent with card transactions
AMTb	Amount spent with card transactions at market type b
INC	Income (0=no,1=low,2=middle,3=high)
PosFreq	Number of pos terminal transactions
NetFreq	Number of transactions on internet
AMTb	Amount spent at type
Have	Already have credit card
Had	Used to have 6 months before the end of training data but not do not have in the last month
Justhave	Obtained credit card in the past 6 months
Longhave	Have credit card for at least 6 months
Lost	Used to have during the 6 months prior the end of training data but not do not have in the last month
Age1	Age below 35
Age2	Age from 36 to 65
Age3	Age above 65
GEN	1 if male, 0 if female
WH	1 if wealthy, 0 otherwise
CIT	1 if the user location is city
VIL	1 if the user location is village
CAP	1 if the user location is capital
CARDS	Amount spent weighted by the ratio showing the percentage of the pos' turnover that was originated from credit cards
POSSD	Standard deviation of the geo x of pos transactions
CLOSE	Number of transactions near to the airport

References

1. https://github.com/dmlc/xgboost
2. https://dms.sztaki.hu/ecml-pkkd-2016/#/app/home
3. R. Kumar, M. Mahdian, B. Pang, A. Tomkins, S Vassilvitskii: Driven by Food: Modeling Geographic Choice. Google. (2015)