Credit Card Usage Prediction

Tamas Veress, 2016 Sep

Introduction

The objective of this challenge was to develop models that describe potential credit card users' behavior to support credit card upselling campaigns

Tasks

- Predict the users' preferred bank branch
- Predict which users are more likely to apply for credit card

Data sets

- User data: age and income group, gender, credit card ownership and application history, address geo info with type of settlement
- Activity time series: credit and debit card transactions with amount, time, location and market category, shop and user IDs
- Bank branch geo info

Method

Branch Visit Prediction

Features

- Distance related: distance from home and median shopping location, rank of distances, cross product of rank and distance
- Location specific: city/town/village
- User specific: gender, age, income group
- Local popularity: how frequently the branch is chosen by users at the same geo location where geo X and Y are rounded to kilometers
- Global popularity: mean prediction error for each branch gathered from a separate training set

Modelling

- Simplify the task to a binary classification problem
- Reduced the user/branch space with selecting only the 5 closest branches the home address and 5 closest to the median card transaction location
- Used XGBoost with area under the curve as evaluation metric for both tasks

Upselling Prediction

New features

- Frequency of transactions made online and via POS terminal
- Amount of transactions made: transformed small, medium, high to 1, 10 and 100 respectively
- Expanded credit card ownership features with separating those who just obtained credit card from those who has one for at least 6 months

Results

Branch Visit Prediction

Figure 1 shows the progress of the branch visit model performance. We separated 50% of the training data on which we measured AUC, Pearson correlation and cosine – evaluation metrics of Task 1

- Models presented:
- Model 1: only distances and rank of distances
- Model 2: added cross product of rank and distance
- Model 3: added settlement type of user address
- Model 4: added user age and gender
- Model 5: added local popularity
- Model 6: added global popularity

Figure 2 presents the feature importance of Model 6.

On Figure 3 we can see the relative importance of features of the blended model. EXCESS denotes the global popularity feature while FREQ, TOP, RAND represent the local popularity variables.

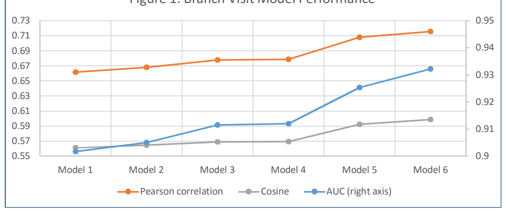

Figure 1: Branch Visit Model Performance

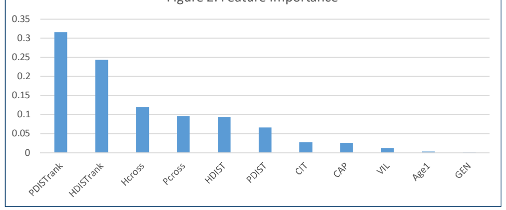

Figure 2: Feature Importance

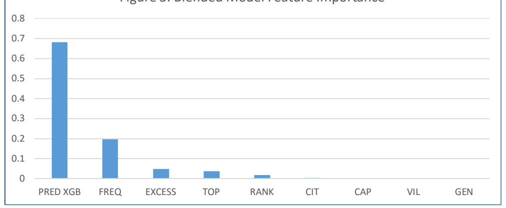

Figure 3: Blended Model Feature Importance

Upselling Prediction

The progress of the upselling models are shown on Figure 4. All scores measured on the entire training user set but data is shifted with 6 months: 2014 H2 features and 2015 H1 card application as target variable

We benchmarked the following model versions to see the added value of processing the transaction data:

- Model 1 includes only features derived from user table: gender, age, income, type of settlement, credit card ownership
- Model 2 only included transaction related features: spending amount and frequency
- Model 3 combined features of Model 1 and 2

Figure 5 displays the feature importance of Model 3

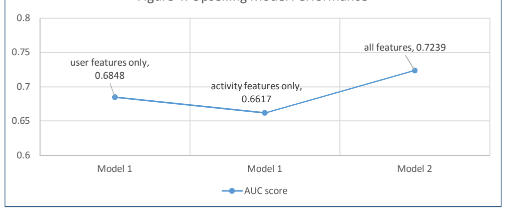

Figure 4: Upselling Model Performance

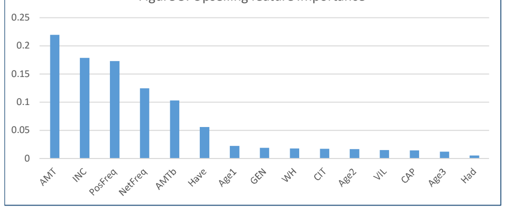

Figure 5: Upselling feature Importance

Conclusion

Branch Visit Model

- Distance from home and card usage location strongly impact the choice of bank branch visited. Home distance is more important when only distance features used
- Rank is stronger feature compared to distance similar to previous studies on geographical choice as it captures the competitive nature of the choice
- Type of settlement (capital/city/village) and user specific features (gender, age) provide further marginal improvements in the prediction accuracy
- Local popularity features add significant gains to our scores while global popularity has relatively little value
- Scores on our local test set suggested that the optimal 'nround' for cosine score is higher than the one for AUC thus using AUC in XGBoost might result underfitting

Upselling Model

- Income group and credit card ownership features are the most important variables gathered from the user data
- Spending amount and frequency features derived from the activity data set significantly improve our prediction
- User gender, age and location type has little impact on credit card application

References

- XGBoost package: https://github.com/dmlc/xgboost
- Tuning XGBoost : https://www.analyticsvidhya.com/blog/2016/01/xgboost-algorithm-easy-steps/
- Bank Card Usage Challenge homepage: https://dms.sztaki.hu/ecml-pkkd-2016/#/app/home
- R. Kumar, M. Mahdian, B. Pang, A. Tomkins, S Vassilvitskii: Driven by Food: Modeling Geographic Choice. Google. (2015)