Machine Learning Engineer Nanodegree - Capstone Proposal

An Xia
December 16, 2018

[Kaggle Competition] Elo Merchant Category Recommendation: Understand Customer Loyalty Using Transactions Data

Domain Background

Elo is a payment solutions company in Brazil, who provides various credit card options to consumers and partners with merchants to offer promotions to cardholders. Elo adds value to cardholders and merchants when cardholders get discounts they highly value and end up purchasing goods and services from the relevant merchants. Therefore, it's highly important to provide targeted promotions to the most relevant group of cardholders. One way to differentiate between high value cardholders vs. the rest is to use loyalty scores, and Elo has called for the data science community on Kaggle to help predict customer loyalty scores given historical credit card transactions data.

Predicting customer loyalty has been a critical topic in marketing for decades. Previous research in this field has successfully applied machine learning techniques such as regression models and neural networks to find the relationship between historical purchases, customer satisfaction, perceived value, and customer loyalty (Buckinx, Verstraeten, & Poel, 2007; Ansari & Riasi, 2016). This project will explore the relationship between historical transactions and customer loyalty using Elo's internal data.

Problem Statement

The goal of this project is to predict cardholder (identified via card_id) loyalty score given historical credit card transactions. Elo has provided training datasets that contain information on 202K cardholders, their loyalty score, and up to 5 months of transactions data. It is unclear how Elo came up with the customer loyalty scores, but examining the dataset shows that customer loyalty score is a numeric value centered around 0 with outliers on the lower end (Fig. 1). Elo has also provided a test dataset that contain features for 124K cardholders, but did not contain any info on their loyalty scores. The goal is to use the historical transactions dataset and training dataset to build a model to predict loyalty score for cardholders in the test dataset.

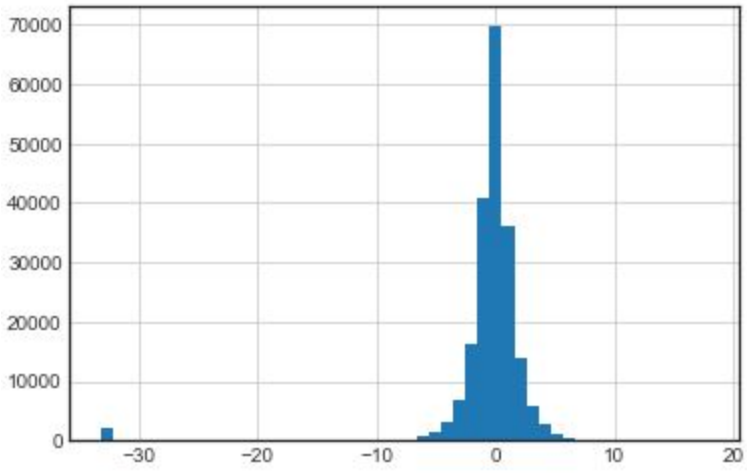

Fig 1. Distribution of customer loyalty score in training data

Datasets and Inputs
Elo has provided the following datasets as inputs on Kaggle (https://www.kaggle.com/c/elo-merchant-category-recommendation/data):

- train.csv - a training set that contains card_id features and loyalty_score for each card_id
- test.csv - a test set that contains card_id features without loyalty_score
- historical_transactions.csv - up to 3 months' worth of historical transactions for each card_id
- new_merchant_transactions.csv - two months' worth of data not presented in historical_transactions
- merchants.csv - additional information about all merchants

Number of observations in each dataset:

- Training data: 201,917
- Test data: 123,623
- Historical transactions: 29,112,361
- New transactions: 1,963,031
- Merchants: 334,696

Solution Statement
Since the desired output is cardholder level loyalty score (continuous numeric value), linear and non-linear regression models are good candidates for this problem. The closer the prediction is to the actual loyalty score, the better the model.

Benchmark Model
A naive benchmark model can be built by segmenting cardholders by buckets of total purchase amounts and assign a loyalty score to each bucket. The training data can be first split up to a train set and a intermediary test set (different from the test data provided by Elo), with both sets containing the true loyalty scores. The train set can then be used to assign purchase amount buckets and their associated scores. Then, the buckets and scores can be applied to the intermediary test set and evaluated against the intermediary test set's true loyalty score. The distance between the predicted scores and true scores on the intermediary test set can be used as a benchmark.

Evaluation Metric
Elo has provided a test dataset, with hidden loyalty scores, serving as the "ground truth". The goal is to minimize the Root Mean Squared Error (RMSE) between the true loyalty score and predicted loyalty score, which takes the root of the average squared difference between the predicted loyalty score for each card_id and the actual loyalty score for the same card_id. The lower the score, the better the prediction. Competition participants will upload the predicted loyalty score for cardholders in the test dataset and Elo will return the RMSE between the submitted score and the true scores.

Project Design

To work towards a solution, the first step is to identify what historical transactions information and what cardholder features are most impactful on the loyalty score. A good starting hypothesis is that high spenders, frequent spenders, spenders that don't get transactions declined, buyers of certain goodies, cardholders located in certain geo locations, longer-term cardholders, etc. are more likely to have higher customer loyalty scores.

With that in mind, the next step is to explore distributions of available features and perform feature engineering. After the feature engineering stage, the following step is to empirically evaluate which features are highly relevant for the model. One technique for feature selection is to use a Decision Tree Regressor to understand feature importances of various features.

Once feature selection is complete, the next step is to select appropriate models to fit the data and run the prediction. Regression models are good options, since the desired output is a continuous numeric variable. Before fitting the models, the training dataset needs to be split into train set and test set for model evaluation based on RMSE. Once a model is selected, grid search can be applied to further tune the model to improve performance.

References

Buckinx, W., Verstraeten, G., & Poel, D. V. (2007). Predicting customer loyalty using the internal transactional database. Expert Systems with Applications,32(1), 125-134. doi:10.1016/j.eswa.2005.11.004

Ansari, A., & Riasi, A. (2016). Modelling and evaluating customer loyalty using neural networks: Evidence from startup insurance companies. Future Business Journal, 2(1), 15-30. doi:10.1016/j.fbj.2016.04.001