ELO Merchant Category Recommendation Project

URL: https://www.kaggle.com/c/elo-merchant-category-recommendation
Datasets: https://www.kaggle.com/c/elo-merchant-category-recommendation/data

Problem Statement

Background:
ELO is one of the largest payment brands in Brazil. The company built partnership with merchants so that it could offer promotions and discounts to its cardholders. Targeted promotions would also help ELO reach out to its most prospective customers with comparatively lower promotional expenses. Promotions are typically beneficial both for the customers as well as the merchants. The promotions are beneficial for customers because (1) they get discount and (2) they don't need to search hard to find merchants that offer discounts. Similarly, it is advantageous for the merchants because (1) they attract more customers into their establishments and (2) get increased revenue for repeat customers.

Goal & Objective:
ELO already figured out the important aspects and preferences in their customers' lifecycle. However, these insights were broadly at an overall level. ELO would like to derive a personalized loyalty score for its customers that will signify how likely they might accept a promotional offer from ELO involving local merchants. Thus, ELO could selectively target audience who are most likely to perform transactions with ELO or visit local merchants after receiving the promotion.

The target column in the training dataset helps ELO achieve this exact goal. The higher the target value was, the higher was the promotion worthiness of the customer. I was interested to build a Machine Learning model that could tell ELO about the customer characteristics that were typically associated with a customer with high target score. My project threw insight on the following questions that remained unanswered.

- Did the customers enjoy their experience?
    - Did the merchants see repeat business?
- What was the level of personalization that can be used for prediction in the model.
- What were the individual customer characteristics associated with promotion worthiness of a customer.
    - The project considered individual traits of customers and the insights from their purchase history. What played a leading role in determining promotion worthiness: (1) Individual traits or (2) Insights from purchase history
    - From Individual traits and Insights from purchase history, what were the best predictors to predict a high target score. A list of few features(not all) in these two categories were:
        - Individual Traits: feature_1, feature_2 and feature_3
        - Insights from Purchase History: Metrics on purchase amount, transaction recency, transaction history, payment logistcs etc.
- Using the data on the cardmembers, transactions and the merchants, could we build a predictive model that was able to accurately predict the target loyalty score for a customer.

Customer/ Target Audience:
The solution was developed for the contest on Kaggle titled 'ELO Merchant Category Recommendation'. Using the loyalty score, ELO would be able to send targeted promotions to the customers with highest prospect.

Nonetheless, the solution and approach can be easily extended to any other business with following goals

- Understand customers' loyalty
    - Can I express customer loyalty through any tangible measures.
    - What factors are typically associated with highly loyal customers
    - Given a customer's characteristics, can I predict the individual's loyalty.
- Incorporate personalized features in models to derive individual customer loyalty
    - Can I use date on customers' prior transactions, their individual traits and purchase habits to build a model to accurately predict customer loyalty.

Data Wrangling

There were 5 datasets associated with the project. The following paragraphs explain how I imported, cleaned and performed data wrangling on the following datasets.

Before initiating the data wrangling activity, I looked at the dictionary to find any hint regarding the reference date on which the current model was run. By analyzing the fields month_lag and purchase_date in historical transactions dataset, it was evident that the reference date was in February 2018. For this project, I considered February 1st, 2018 as the reference date/model run date.

Training Datasets

- First, I imported the training dataset. The first active month field was parsed as date during the import activity.
- There were no null values observed in the dataset.
- I included a calculated field 'elapsed active days' to the dataset. It was calculated as the difference between the reference date and the first active month.
- There are a few categorical features in the dataset: feature_1, feature_2 and feature_3. They were converted to numeric features through one hot vectorization.
- Finally, I dropped the first active month field before modeling.

Merchants Dataset

- Next, I imported the Merchants dataset for analysis.
- I converted the categorical features (most recent sales range, most recent purchase range columns) to numeric columns through one hot encoding.
- The fields category_1 and category_4 were boolean columns with Y/N values. I converted these fields to numeric columns with 1/0 values.
- The category_2 column had a few nulls. I replaced the nulls values in the field with the median of category_2 field. The field was converted to a series of numeric columns through one hot vectorization.

Historical Transactions Dataset

- I imported the historical transactions dataset for analysis. While import, I parsed the purchase date field as a date.
- The dataset had a few boolean columns (e.g. authorized_flag and category_1). I transformed these columns into numeric columns with 0/1 values.
- Next, I derived a field purchased elapsed days which is essentially the difference between the model run date and the purchase date. The purchase date field was deleted from the dataset before modeling.
- There were a few categorical features (e.g. category_2 and category_3) in the dataset. I converted them to numeric features through one hot vectorization.
- **Aggregation:** As the historical transactions dataset was at transaction level, I aggregated it at card_id level to get aggregated metrics. I joined the training dataset to aggregated historical transactions dataset so that insights could be found from the combined dataset.

New Merchant Transactions

Next, I employed the same data wrangling functions on the new merchants transactions dataset as well

Test Dataset

The data wrangling steps applied on the training dataset, was also applied to the test dataset.

Preparation of Combined Dataset

There were a few unique challenges observed in the project because of (1) availability of multiple datasets at different granularity and (2) high volume of transaction data. We overcome the challenges by following methods:

Aggregation of Datasets: The granularity of the datasets in the project were as follows

Datasets	Granularity
Training,Test	card_id
Historical Transactions, New Merchant Transactions	card_id, transaction timestamp
Merchants	merchant_id

Table 1: Granularity of datasets

I prepared the combined train and test dataset as follows

- First, I joined the historical transactions with the merchants dataset at merchant_id level to fetch the merchant specific attributes. The new merchant transactions dataset was also joined with the merchants dataset for similar insight.
- Next I aggregated the historical transactions + merchants dataset at card_id level to derive the metrics of interest at aggregate level. Then, the train dataset was combined with the aggregated historical transactions dataset on card_id. Similarly, I aggregated the new merchant transactions + merchants dataset at card_id level and combined it with the training dataset. Both of these

combined datasets were joined to form the final dataset which was subsequently used for model training.
- I repeated the exact same procedure using the test dataset to obtain the combined test dataset. This will be used for model validation.

Decomposition of Datasets (Divide and Conquer)

The historical transactions dataset contained 29 million transaction and the new merchants dataset had 1.9 million transactions. In order to reduce the processing time of the datasets, I added an additional feature in the training dataset that divided the target dataset into 3 equal sets with low, medium and high target respectively. The card-ids in these divided datasets were joined with historical transactions dataset and new merchants dataset and were aggregated. Finally. the aggregated datasets for card_ids from low, medium and high target were concatenated to form the final aggregated dataset. I processed these datasets and saved the intermediate aggregate datasets that helped to reduce the end-to-end processing time during data wrangling activity.

As a next step, I conducted exploratory data analysis and inferential statistics for detailed analysis of the datasets. Based on the insights, I might perform additional data wrangling steps on the datasets.

Exploratory Analysis

In the exploratory analysis phase, I performed visual exploratory data analysis and statistical analysis.

Training Dataset

I initiated my exploration with the training dataset because this dataset contains the variable of interest (target). My observations were as follows

- Target Variable
    - The histogram of the target variable revealed that the target followed a normal distribution with a high peak and narrow tails. So, a higher percentage of the target variable was near its mean and the distribution dropped rapidly as the target increased or decreased in either direction.
    - Then, I visualized the target variable in a box plot to check for outliers. Few extreme outliers were found with a target value less than -30.

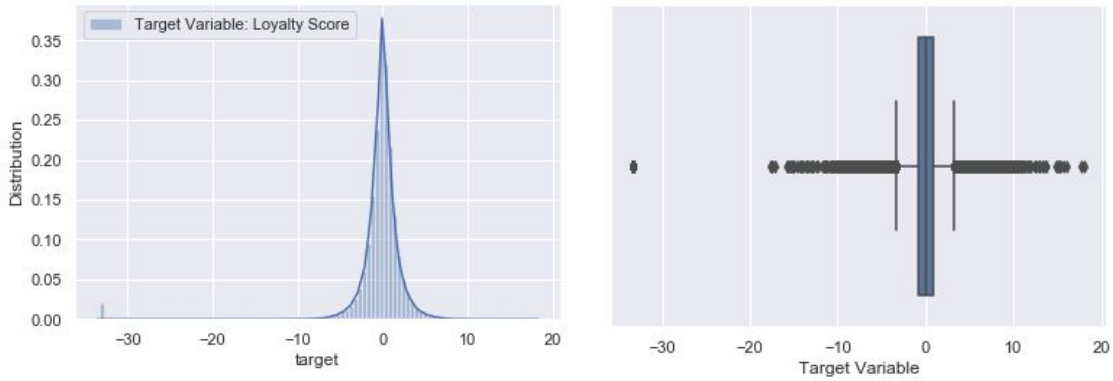
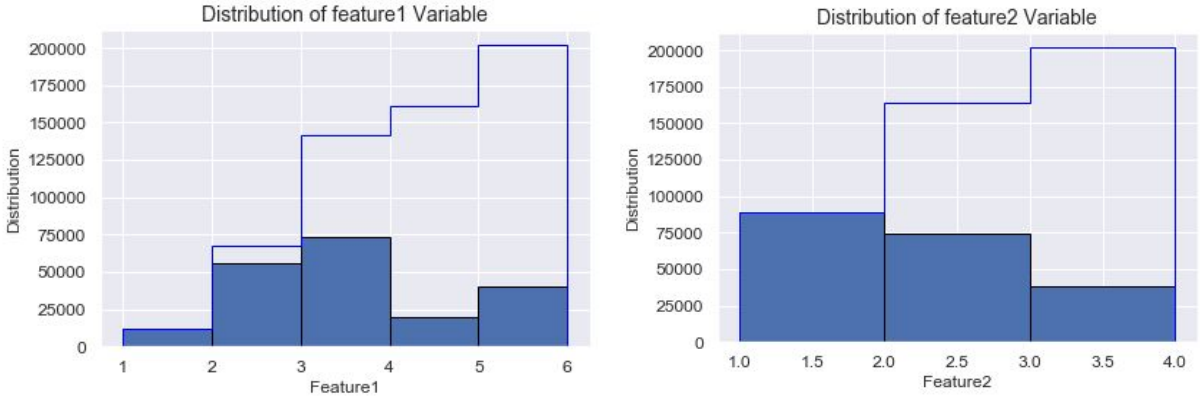


Figure 1: Distribution of Target Variable (Histogram and box plot)

- Next, I visualized the 3 categorical features (feature_1, feature_2 and feature_3) in the training dataset. As these features are at card_id granularity, these seem to be characteristics related to the cardholders. The following characteristics were observed.
    - feature_1: Most of the cardholders had a value 2 or 3 for feature_1 variable.
    - feature_2: The feature_2 variable was inversely proportional to the distribution of card holders. Highest number of cardholders were associated with feature_2 value 1. The frequency dropped as the value of feature_2 increased.
    - The boxplots of featue_1 and feature_2 variables revealed that the median of target score did not differ much with different values in feauture_1 and feature_2 variables. Based on the visual exploration, neither the feature_1 nor the feature 2 had significant influence on target variable.



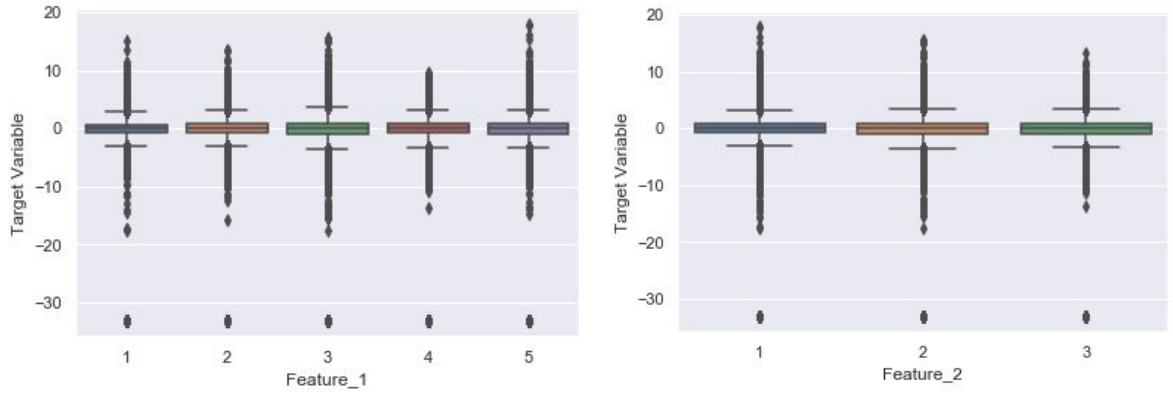


Figure 2: Distribution of feature_1 and feature_2 variables

- feature_3: The feature_3 variable was a boolean feature with value 0 or 1. There were slightly higher number of cardholders with value 1 compared with the number of cardholders with value zero. The box plot revealed that the feature_3 variable did not have noticeable influence on target variable.

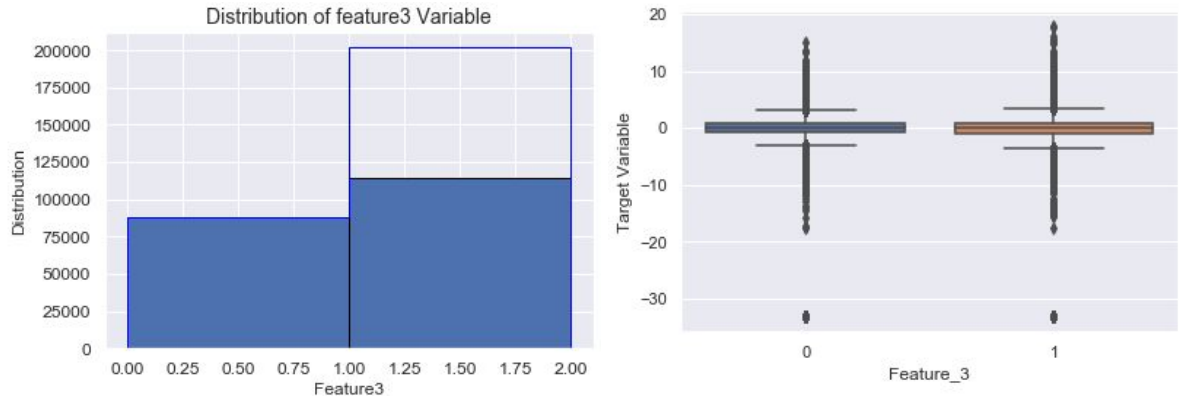


Figure 3: Distribution of feature_3 variable

The plot of elapsed days and the target variable assumed a funnel shape hinting that the target score leaned more towards zero as the members' tenure increased. The elapsed day is the difference between the model run date and the first active month.

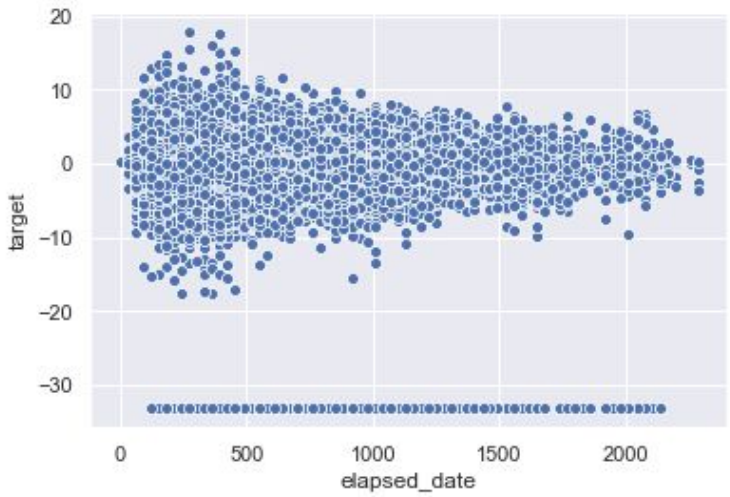


Figure 4: (i) The scatter plot (elapsed days, target)

Next, I extracted the month from first active month and plotted the box plot of the target variable with it. As we observed, each month had the outliers < -30 observed earlier. Next, I created the above plot by excluding the outliers in training dataset. As noticed, now the variations in target score are comparatively more clear. Based on the box plots, the start_month variable did not seem to have a significant influence on target median score.

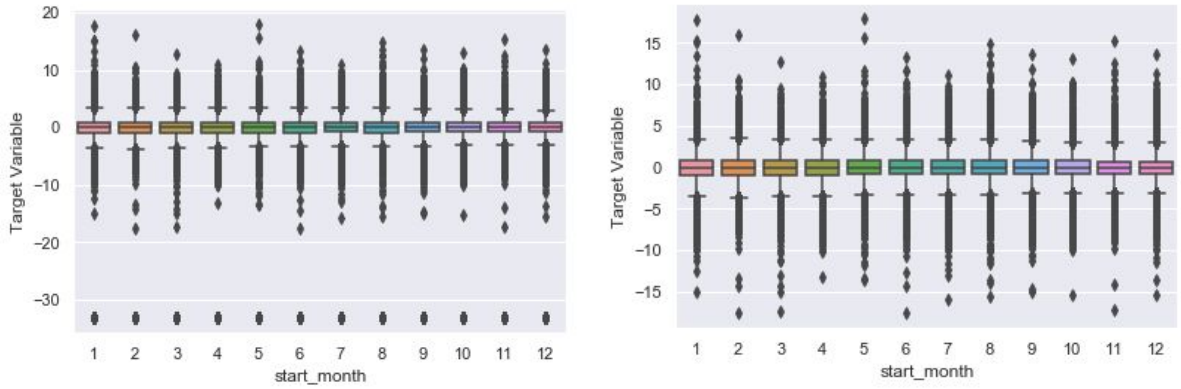


Figure 5: Box plot (start_month and target) after outliers (target <=30) were removed

Next, I was interested to discover any interesting pattern on the outliers dataset. I created a separate outliers dataset that only consisted the cardIDs that were less than -30 and combined it with information from historical transactions, new merchants and merchants dataset. It was observed that the city with ID 333 had the highest purchase amount and city_id -1 had the highest transactions. Based on the exact sum of purchase amount and the the transaction count, it was observed that 1% of total cards were in outliers dataset whereas (i) the purchase amount sum was 2% of total purchase amounts and (ii) the number of transactions were 7% of total transactions.

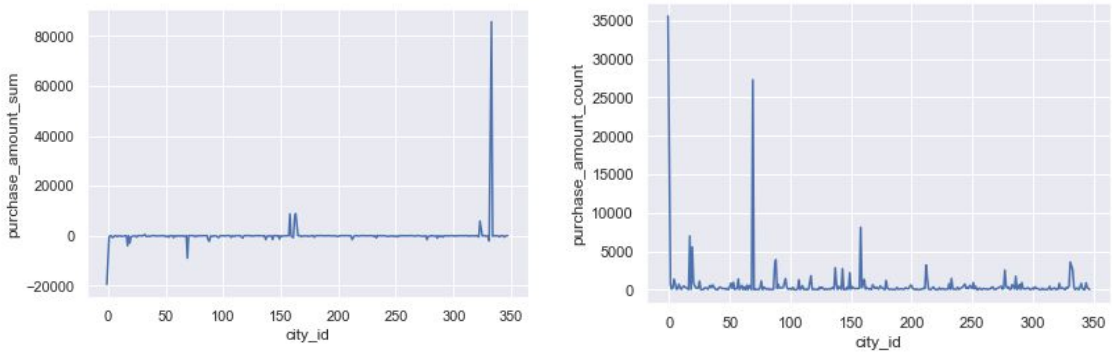


Figure 6: City ID vs Purchase Amount Sum and Count (For Outliers Only dataset).

Outliers were excluded from the dataset for subsequent steps in exploratory and visual data analysis. I was interested to explore how the features and transactions for a cardholder with high target score and low target score differed. For further analysis, I grabbed (1) the top 10% and (2) the bottom 10% observations in training dataset based on target score. I joined the datasets with historical transactions and new merchant transactions and aggregated the transactions in order to analyze any patterns. Two cardIDs(with same first active month) were selected at random from the top 10% and the bottom 10% dataset respectively. The same first active month was used for both the cards in order to nullify any impact from (1) purchase transactions based on seasonality and (2) membership tenure.

Card ID	Target	Dataset
C_ID_643ddaea8c	3.459307	Top 10% Target Score
C_ID_24b7e16c4f	-2.788379	Bottom 10% Target Score

Table 1: The two card_ids with high and low target score selected at random for analysis

The aggregated measures for the above two cards are mentioned as follows:

	Card_ID: C_ID_643ddaea8c (High Target Score)	Card_ID: C_ID_24b7e16c4f (Low Target Score)
Target Score	3.459307	-2.788379
Unique Cities Count (H)	4	9
Unique Merchants Count(H)	45	21
Unique Merchant Group ID (H)	34	15
Purchase Amount Sum (H)	-75.41	-24.44
Elapsed Days Range	352	243
Numerical_1 Sum	2369	43.5
Numerical_2 Sum	2316	43.4

Abbreviations: H: Historical Transactions, N: New Merchant Transactions

Table 2: The aggregated measures for the two card_ids selected at random

I formed the following hypothesis from my observations that I will revalidate in the modeling stage.

Hypothesis

- The cardID with high target score was associated with higher
    - Count of Unique Merchant IDs
    - Absolute value of historical purchase amount sum
    - Historical elapsed time range
    - Count of Unique Merchant Group IDs
    - Count of Unique Historical Merchant Categories
    - Sum of Numerical_1 quantity
    - Sum of Numerical_2 quantity
- Similarly the card with higher target score was associated with lower unique City IDs.

Next, I analyzed the merchants, historical_transactions and new_merchant_transactions dataset.

Merchants Dataset

- I found a linear relationship between the features numerical_1 and numerical_2 with a pearson correlation coefficient 0.999. As the features are highly correlated with each other, I took out numerical_2 feature from the dataset to avoid multicollinearity.

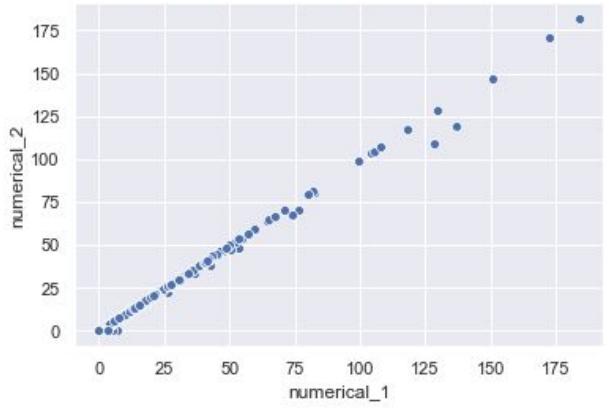


Figure 7: Scatter plot (numerical_1, numerical_2) showed approximately perfect linear relationship

- Next, I analyzed the impact of transactions frequency on the features: (i) most recent sales range and (ii) most recent purchases range. From below histogram, it is evident that most of the transactions were associated with merchants that had most recent sales range 5 and most recent purchase range 5 respectively.

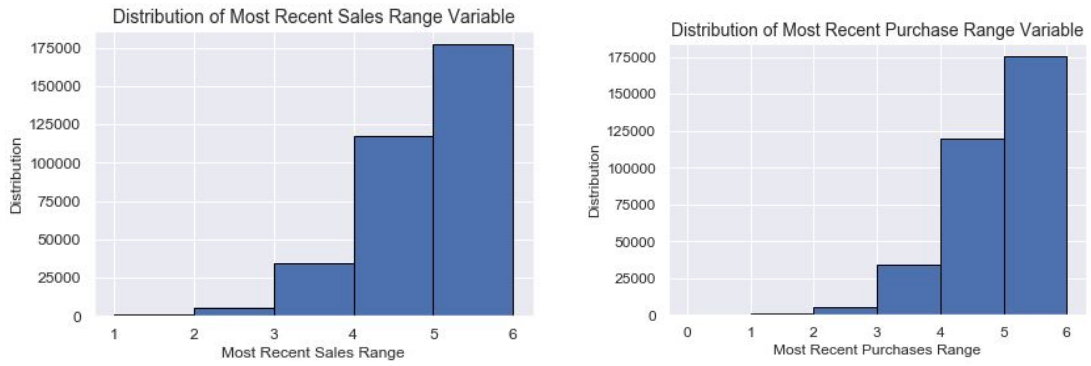


Figure 8: Distribution of Merchants in Most Recent Sales Range and Most Recent Purchase Range

Historical Transactions Dataset

Next, the purchase amount field was aggregated based on city_id and few interesting insights were discovered.

- The number of purchase transactions and the sum of purchase amount were the highest for City 69. In fact, the total purchase amount of city 69 seemed to be higher than all the cities combined. It can be implied that city 69 may be an important city in Brazil with high number of ELO card usage.
- The purchase count associated with city -1 was very high, however the sum of purchase amount with city -1 was very low. This means these may be small cities where low value transactions constitute a higher proportion from all transactions.
- In order to capture these city specific insights, I included three variables to the dataset that signified whether the city_id was -1, 69 or 333 respectively.

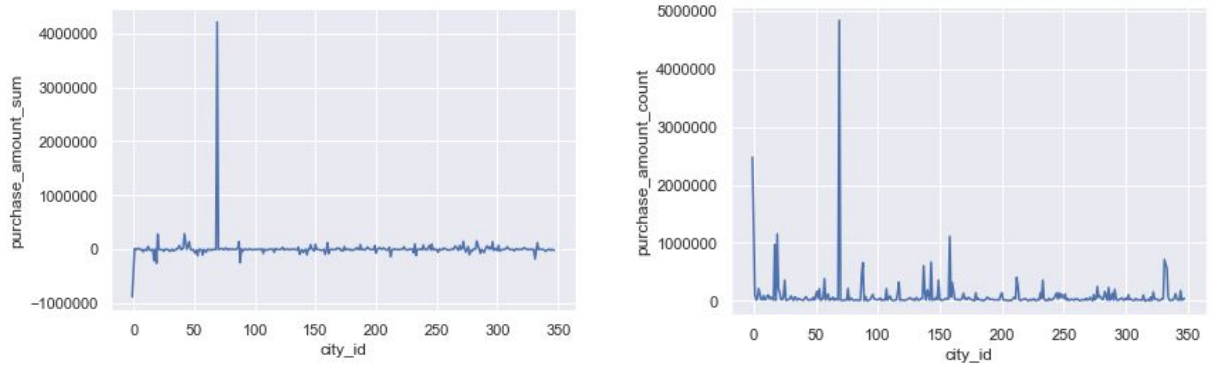


Figure 9: Distribution of total purchase amount and total purchase count by city_id

I built the aggregated datasets following the two methods ((i) Aggregation, (ii) Decomposition) described in the section: Preparation of Combined Dataset above. I had the following observations from the aggregated dataset.

Observations from the aggregated dataset

- Most cardholders with a higher sum of absolute normalized purchase amounts from new merchants dataset leaned towards the target score.

Figure 10: Scatter Plot (New Merchants Purchase Amount Sum (Absolute) vs Target variable)

- Similarly, I also observed that the target value more towards 0 as the transaction count (both historical as well as new merchant) increased.

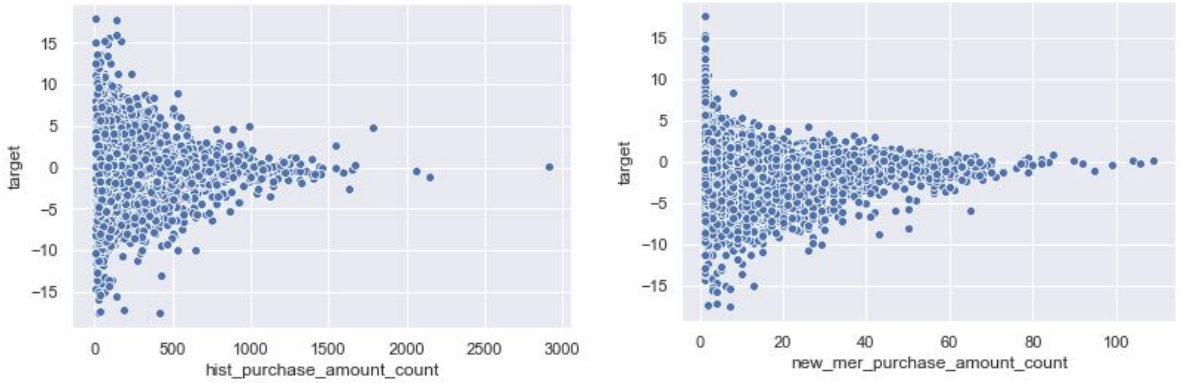


Figure 11: Influence of Purchase Transaction Counts on Target variable

- I also had an observation that higher median purchase amounts in historical transactions as well as the new merchant transactions dataset were close to zero.

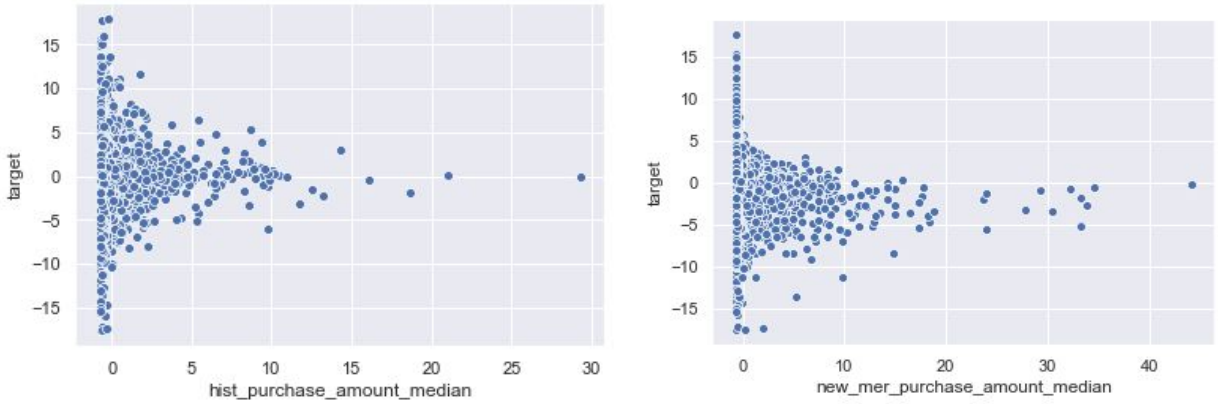


Figure 12: Influence of purchase transaction median on target variable

- The cardmembers with high unique cities of historical transactions mostly had a score close to zero. I also observed a similar trend between the unique cities in new merchant transactions and the target variable.

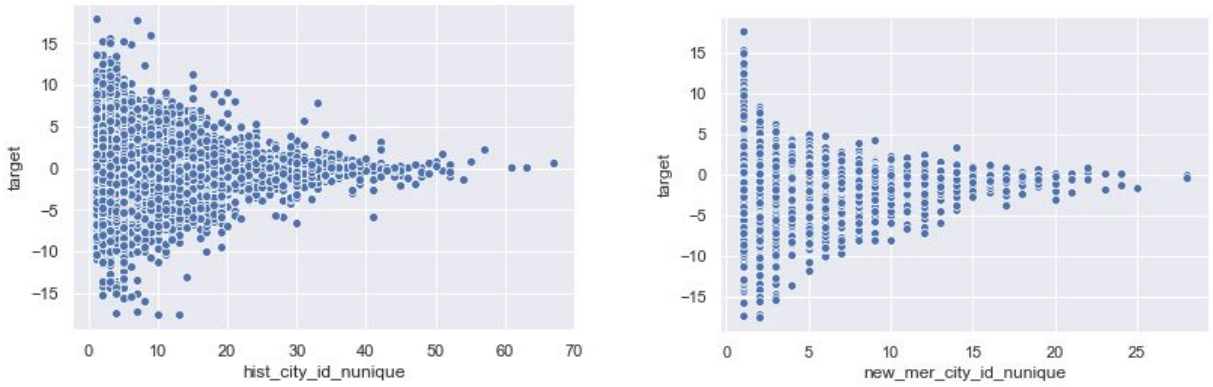


Figure 13: Influence of unique number of cities (transactions) on target

- Next, I found the cardmembers with high purchase date range typically had a target score away from zero. On the other hand, the cardmembers with a low purchase history had a score close to zero.

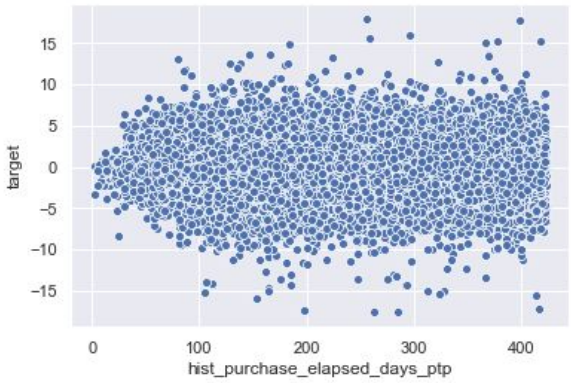


Figure 14: Influence of date range of transactions on target variable

As I am at the end of the Exploratory Data Analysis stage, Let's revisit the questions we were trying to answer at the initiation of the project.

- **Did the customers enjoy their experience? Did the merchants see repeat business?**
    - Yes, Most customers revisited the merchants they already went to.
- **What was the level of personalization that can be used for prediction in the model ?**
    - There is good amount of personalization involved because of availability of following features in dayase:
        - Feature variables associated with cardholders (feature_1, feature_2 and feature_3 variables in training dataset)
        - Length of membership
        - Features associated with cardmembers' purchase transactions (installments, other features, purchase amount, elapsed days since purchase days)
        - Characteristics of merchants at which the cardmembers shop.

- **What were the individual customer characteristics associated with promotion worthiness of a customer ?** Based on the exploratory data analysis, I formed the below **hypothesis**.
    - The following characteristics can be associated with customers with a high target score.
        - High Unique Merchant ID Count, High Unique Merchant Group ID Count
        - High Purchase Amount Sum
        - Higher Elapsed Days Range
        - High Sum of Numerical_1 Quantity
        - Lower Unique Cities of Transaction

- **Using the data on the cardmembers, transactions and the merchants, could we build a predictive model that was able to accurately predict the target loyalty score for a customer.**
    - Yes, I would build a predictive model in subsequent stage to validate my hypothesis and to find any other interesting insights I might have missed in Exploratory Data Analysis.
    - The predictive model would also help find the features with highest influence on the target score.

Modeling

Upon inspection of the training aggregated datasets and the test aggregated datasets, it was evident that many variables had different numeric range. I normalized the training dataset using a standard scaler to eliminate any influence from different variable ranges. The test dataset was passed through the same standard scaler transformer to preserve the same normalization parameters used for the training dataset.

LightGBM Full Model

First, I built a Light Gradient Boosting Machine(GBM) algorithm from all the features to predict the relationship between the aggregated features and the target variable. I chose the light GBM model because of the following reasons.

1. LightGBM is an ensemble boosting algorithm.
2. As LightGBM is based on decision trees, the importance of the features in the model influencing the target can be easily explained.
3. Compared with other accurate algorithms (e.g. extreme boosting algorithm), LightGBM model can be trained very fast.

Though, the LightGBM model implementation has many benefits, it also comes with a challenge. For optimal performance, the hyper-parameters have to be tuned well. I employed CV GridSearch to tune the hyper-parameters and found the following optimal hyper parameters. They were used in building the LightGBM model.

<u>Optimal Hyperparameters</u>

- Learning Rate: 0.01
- Boosting Type: gbdt

- Objective: Regression
- Metric: l2_root
- sub_feature: 0.5
- num_leaves: 30
- min_data: 50
- max_depth: 7

As I was looking for a low RMSE score, I chose l2_root as a metric of choice for regression. In order to avoid overfitting, I implemented a 10 fold cross validation with shuffling to build the final model. The best model RMSE obtained from the cross validation was 3.5762.

Feature Importance and Feature Selection

Then, I visualized the features based on their importance. As hypothesized before, the features built on purchase elapsed time and purchase amount had high influence on the target. My analysis also revealed that the sum of category_1 measure from historical transactions is an important predictor of the target variable.

feature	importance
hist_purchase_elapsed_time_min	1863.8
new_mer_purchase_elapsed_time_min	1637.0
new_mer_purchase_amount_max	1161.0
hist_category_1_x_sum	1017.7
hist_purchase_elapsed_time_ptp	969.9
hist_authorized_flag_mean	945.9
new_mer_purchase_elapsed_time_max	916.4
new_mer_purchase_elapsed_time_mean	878.8
new_mer_purchase_elapsed_time_ptp	874.5
hist_purchase_elapsed_time_mean	846.5

Figure 15: Features with highest influence on target variable

As this model was built on more than 160 features, I decided to build a subsequent model from a reduced set of features (top 40 features based on their importance). This step enabled me to see the influence of the top predictors after eliminating any noise from the other predictors.

LightGBM Reduced Model

I built a Light GBM model from the top 40 features (around 25% of all the features) of the full model. Again, I employed CVGridSearch to tune the hyper-parameters and obtained the

optimal hyperparameters. The optimal hyper-parameters were used to build the model and the best model was obtained through cross validation. The best model from cross validation had a best score 3.5898. Here is a screenshot of the top 15 most important features in the reduced model.

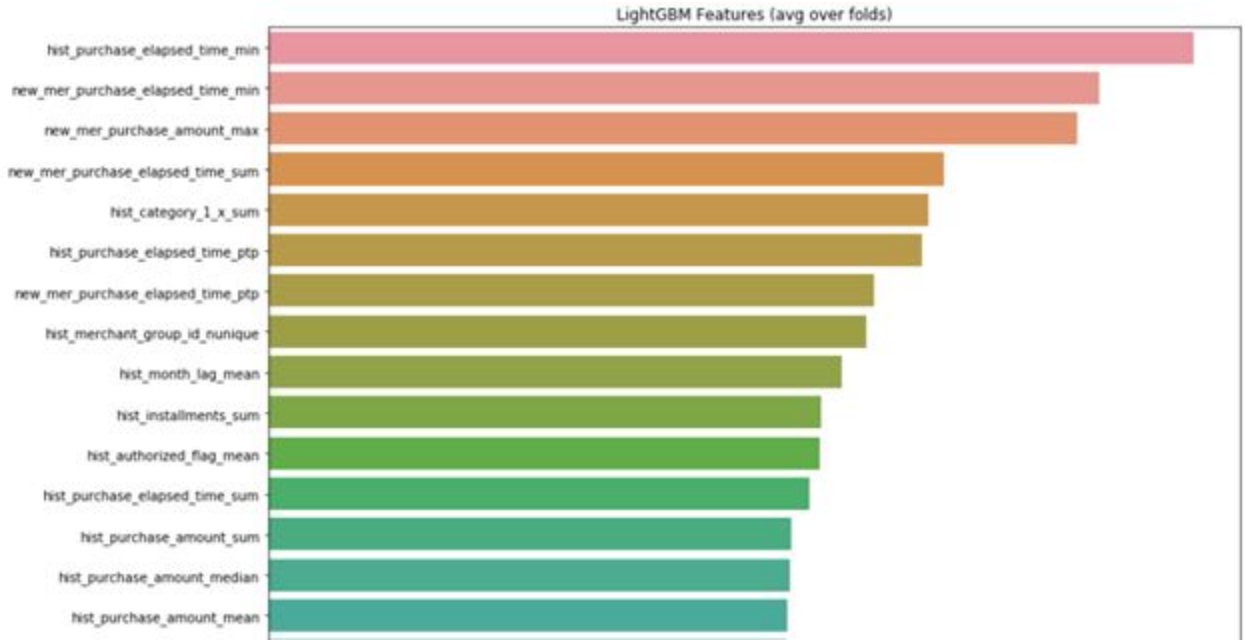


Figure 16: Feature Importance of Reduced Model

Ridge Regression

I decided to build a 2nd model based on ridge regression to check whether it provides a better score. I chose Ridge Regression because the algorithm puts a penalty on features with high values, which in turn helps to reduce fluctuations on RMSE score because of large values in features. I obtained the optimal value of alpha through hyperparameter tuning using CVGridSearch. Using the optimal value of alpha, I built the final ridge regression model.

In subsequent sections, I presented a comparison of model performance of various models and discussed recommendations for ELO how it should selectively target its customers for promotions.

Model Validation

Minimum RMSE (Root MEan Square Error) was the evaluation metric for the ELO Merchant Category Recommendation challenge. I presented below the list of RMSE obtained by all the models I built. As shown in the table, the LightGBM full model had the best RMSE score among the 3 models we evaluated.

Model	Training Dataset (CV Score)	Test Dataset (Submitted)
LightGBM Full	3.5762	3.71266
LightGBM Reduced	3.5898	3.71770
Ridge Regression		3.717

Table 3. RMSE Score for All Models

Recommendations

ELO planned to send targeted promotions to its customers who will most likely use the promotions. Based on insights discovered from the exploratory data analysis and the modeling stage, I had few recommendations for ELO so that it can effectively targets customers with a high target score.

I would recommend ELO to use the machine learning model to predict the target score for customers and then send targeted promotions to customers with high target scores. Typically, the following customer characteristics are great predictors of a high target score.

- **Recency of Transactions:** Customers who purchased in recent past**.**
- **High Aggregated Purchase Amount**: Customers with high aggregated purchase amount
- **High number of old transactions**: Customers with a high number of old transactions
- **Sum of Category_1:** Customers with a high aggregate sum of category_1.
- **Long Data Range of Purchases:** Customers with a longer date range of purchases.
- **High number of Unique Merchants:** Customers with transactions that involved a high number of unique merchants.
- **High Duration Installments**: Customers who purchase items with high duration installments.
- **Transaction Characteristics**: Customers with high transactions associated with category_1(value 1) and category_3 (value B).

My above findings lent support in favor of following parts of the hypothesis

- The following characteristics can be associated with customers with a high target score.
    - High Unique Merchant ID Count
    - High Purchase Amount Sum
    - Higher Elapsed Days Range

However, the following factors in hypothesis did not prove to be a significant predictor for target score.

- High Unique Merchant Group ID Count
- High Sum of Numerical_1 quantity
- Lower Unique Cities of Transactions.

Future Work

My current work used LightGBM and Ridge Regression algorithms to build a predictor to tackle the ELO Merchants Category Recommendation challenge. In future, I would like to implement a neural network-based model as well and build an ensemble out of all models that might have better predictability. There is also scope of using clustering to group merchants based on their demographic and geographic association. Interesting insights on ELO card usage across various geographic entities (states, cities might also improve model performance.

Reference

I am thankful to the participants of ELO Merchant Category Recommendation Kaggle contest who shared public kernels and their invaluable insights on the Kaggle Challenge.