Capstone Project
Data Science Nanodegree
Ayidh Alqahtani

Definition:

Project Overview:

Elo, one of the largest payment brands in Brazil, has built partnerships with merchants in order to offer promotions or discounts to cardholders. But do these promotions work for either the consumer or the merchant? Do customers enjoy their experience? Do merchants see repeat business? Personalization is key.[1]

Problem statement:

The Elo wants to predict the customer loyalty score to reduce expenses on unwanted advertisement campaign which costs a lot of money. This problem is a supervised learning problem and since it's predicting the score (number) , so it is a regression problem. The input data are card_id ,merchant_Id , city_id , Category and other features. The output is going to be the target which is a customer's loyalty score.

Metrics:

The evaluation metrics for this project is going to be Root Mean Squared Error (RMSE).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

where ŷ is the predicted loyalty score for each card_id, and y is the actual loyalty score assigned to a card_id.

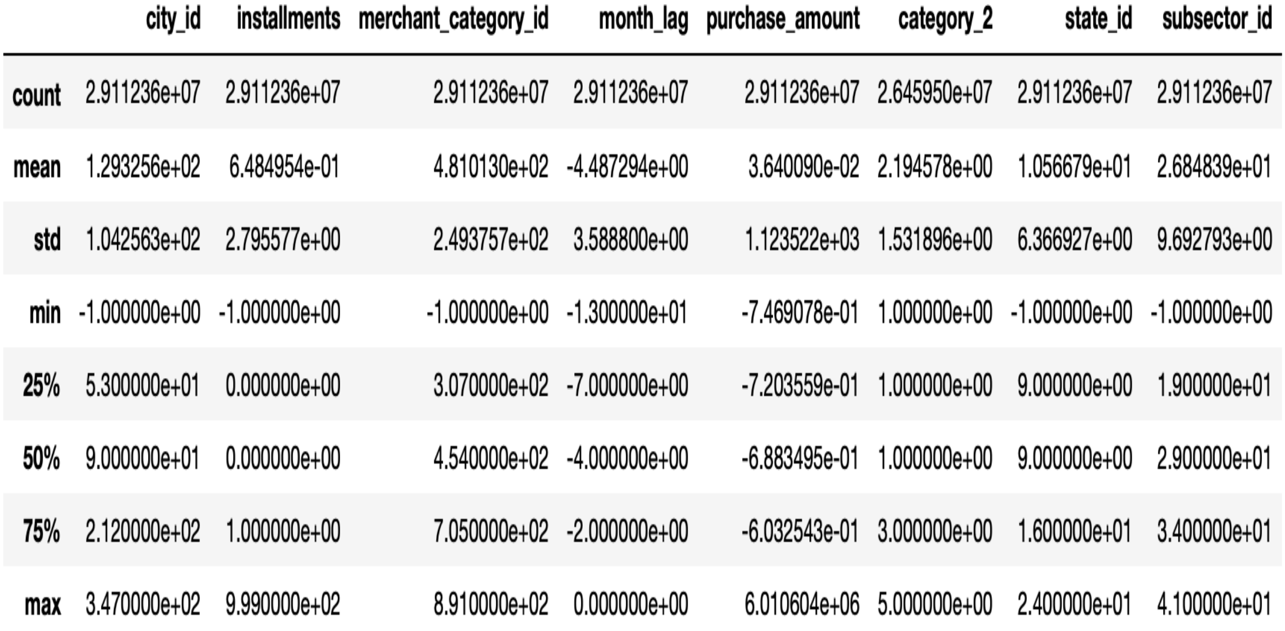
Analysis:

Data Exploration:

There are different files have been provided by the ELO merchant including : historical transactions , merchants , new_merchants ,train data set , test data set which can by joined with other files. The historical transactions contain the transaction for each card id up to 3 months at any provided merchant id. New merchants transactions have the new transactions at the new merchant (merchant_ids that this particular card_id has not yet visited) over a period of two months. merchants.csv contains aggregate information for each merchant_id represented in the data set. [1]
This data set was provided by ELO merchant competition in the Kaggle website and these data set can be found in the below link:
https://www.kaggle.com/c/elo-merchant-category-recommendation/data

historical_transacatios data set description :

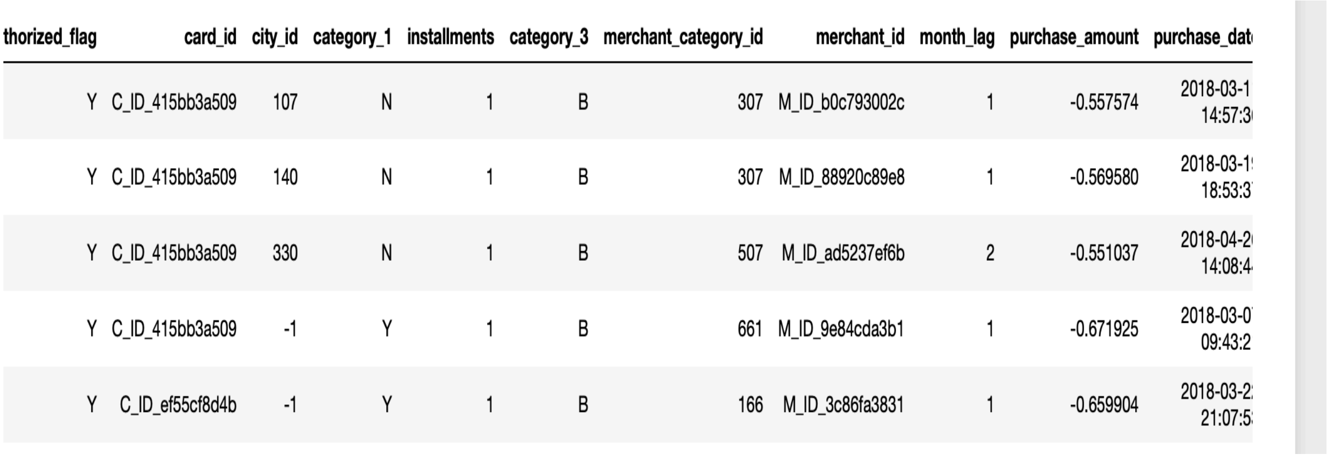
Historical transactions dataset has 29112361 data points with 14 varia bles each.



	city_id	installments	merchant_category_id	month_lag	purchase_amount	category_2	state_id	subsector_id
count	2.911236e+07	2.911236e+07	2.911236e+07	2.911236e+07	2.911236e+07	2.645950e+07	2.911236e+07	2.911236e+07
mean	1.293256e+02	6.484954e-01	4.810130e+02	-4.487294e+00	3.640090e-02	2.194578e+00	1.056679e+01	2.684839e+01
std	1.042563e+02	2.795577e+00	2.493757e+02	3.588800e+00	1.123522e+03	1.531896e+00	6.366927e+00	9.692793e+00
min	-1.000000e+00	-1.000000e+00	-1.000000e+00	-1.300000e+01	-7.469078e-01	1.000000e+00	-1.000000e+00	-1.000000e+00
25%	5.300000e+01	0.000000e+00	3.070000e+02	-7.000000e+00	-7.203559e-01	1.000000e+00	9.000000e+00	1.900000e+01
50%	9.000000e+01	0.000000e+00	4.540000e+02	-4.000000e+00	-6.883495e-01	1.000000e+00	9.000000e+00	2.900000e+01
75%	2.120000e+02	1.000000e+00	7.050000e+02	-2.000000e+00	-6.032543e-01	3.000000e+00	1.600000e+01	3.400000e+01
max	3.470000e+02	9.990000e+02	8.910000e+02	0.000000e+00	6.010604e+06	5.000000e+00	2.400000e+01	4.100000e+01

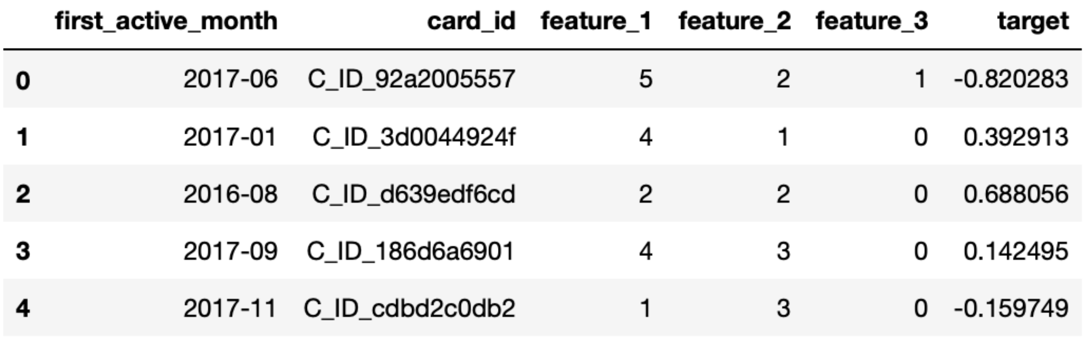
Merchant data set description:

new_merchant_transc dataset has 1963031 data points with 14 variables each.



thorized_flag	card_id	city_id	category_1	installments	category_3	merchant_category_id	merchant_id	month_lag	purchase_amount	purchase_dat
Y	C_ID_415bb3a509	107	N	1	B	307	M_ID_b0c793002c	1	-0.557574	2018-03-1 14:57:3
Y	C_ID_415bb3a509	140	N	1	B	307	M_ID_88920c89e8	1	-0.569580	2018-03-1 18:53:3
Y	C_ID_415bb3a509	330	N	1	B	507	M_ID_ad5237ef6b	2	-0.551037	2018-04-2 14:08:4
Y	C_ID_415bb3a509	-1	Y	1	B	661	M_ID_9e84cda3b1	1	-0.671925	2018-03-0 09:43:2
Y	C_ID_ef55cf8d4b	-1	Y	1	B	166	M_ID_3c86fa3831	1	-0.659904	2018-03-2 21:07:5

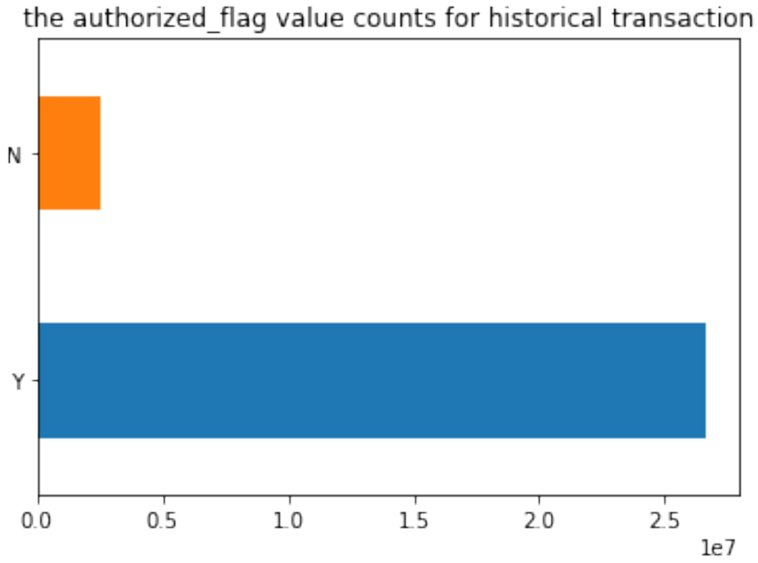
Train data set:

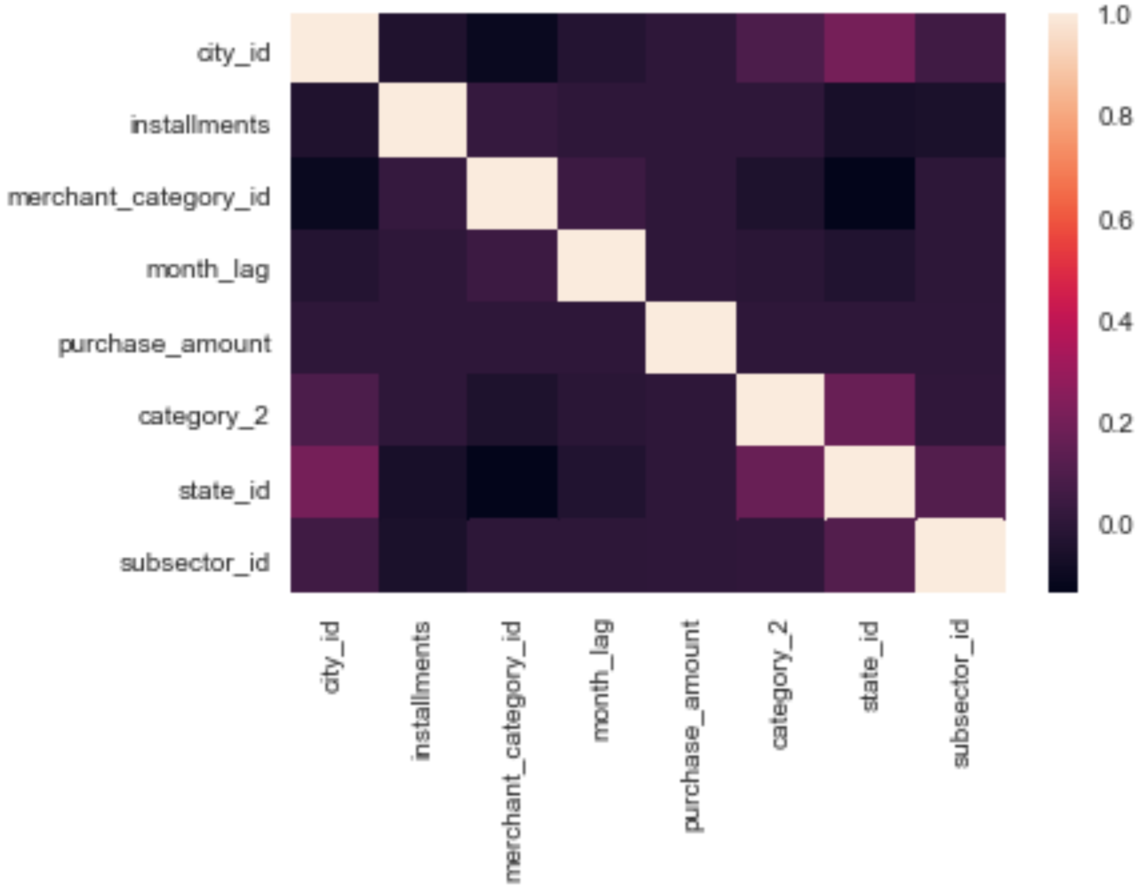


	first_active_month	card_id	feature_1	feature_2	feature_3	target
0	2017-06	C_ID_92a2005557	5	2	1	-0.820283
1	2017-01	C_ID_3d0044924f	4	1	0	0.392913
2	2016-08	C_ID_d639edf6cd	2	2	0	0.688056
3	2017-09	C_ID_186d6a6901	4	3	0	0.142495
4	2017-11	C_ID_cdbd2c0db2	1	3	0	-0.159749

Exploratory Visualizations:

In the below graph shows the distribution of authorized flag for the transaction in the data set.





We have created correlation matrix for the data set, however , it seems there is no strong correlation between these features as can be seen above.

Algorithms and techniques:

The purpose of this project is to predict the customer's loyalty score which is number. So, we are going to build regression model. Furthermore, we will use different algorithms including Linear regression algorithm, Random Forest algorithm, and Gradient Boosting Regressor algorithm.

Benchmark:

We would like to use the Root Mean Squared Error (RMSE) to evaluate our model results. The benchmark model is Stochastic Gradient Descent (SGD)

https://scikit-learn.org/stable/modules/sgd.html#regression

we can compare our model to the benchmark model and we can know how our model performs by doing that.

Methodology:

Data Preprocessing:

Before this step, we have data exploration which gave us insight about the data set and its distribution among each other. We have noticed there are some nulls values for some column like category2 and category 3 which we dropped because we think there are not gonging to affect our model neither the results.

Further, for the merchant id nulls, we filled the nulls with zeros because we think it is a good think to have this feature. We merge two data set with each by card it to take advantage of these features and enrich our model. Now, our data is neat and ready to be used.

Implementation:

In implementation phase, we have gone through different steps as below:

Step 0: Import the data and the necessary libraries.
Step 1: Data Exploration and visualization.
step 2: Data cleansing and preparation.
Step 3: Model selection and evaluate the algorithm.
Step 4: Model tuning.
step 5: Results

In the model selection and the chooses algorithms, we have tested three different algorithms :

- Linear regression algorithm
- Random Forest algorithm
- Gradient Boosting Regressor algorithm

Our evaluation metrics for the model is Root mean squared error and the results as below :

Refinements:

We have done feature engineering by adding some features to the training and we change the value of random_state=500 instead of 300 .

Before this tuning the rsme was 3.81 for the linear regression and we got 3.61 for the same model after doing this tuning

Results:

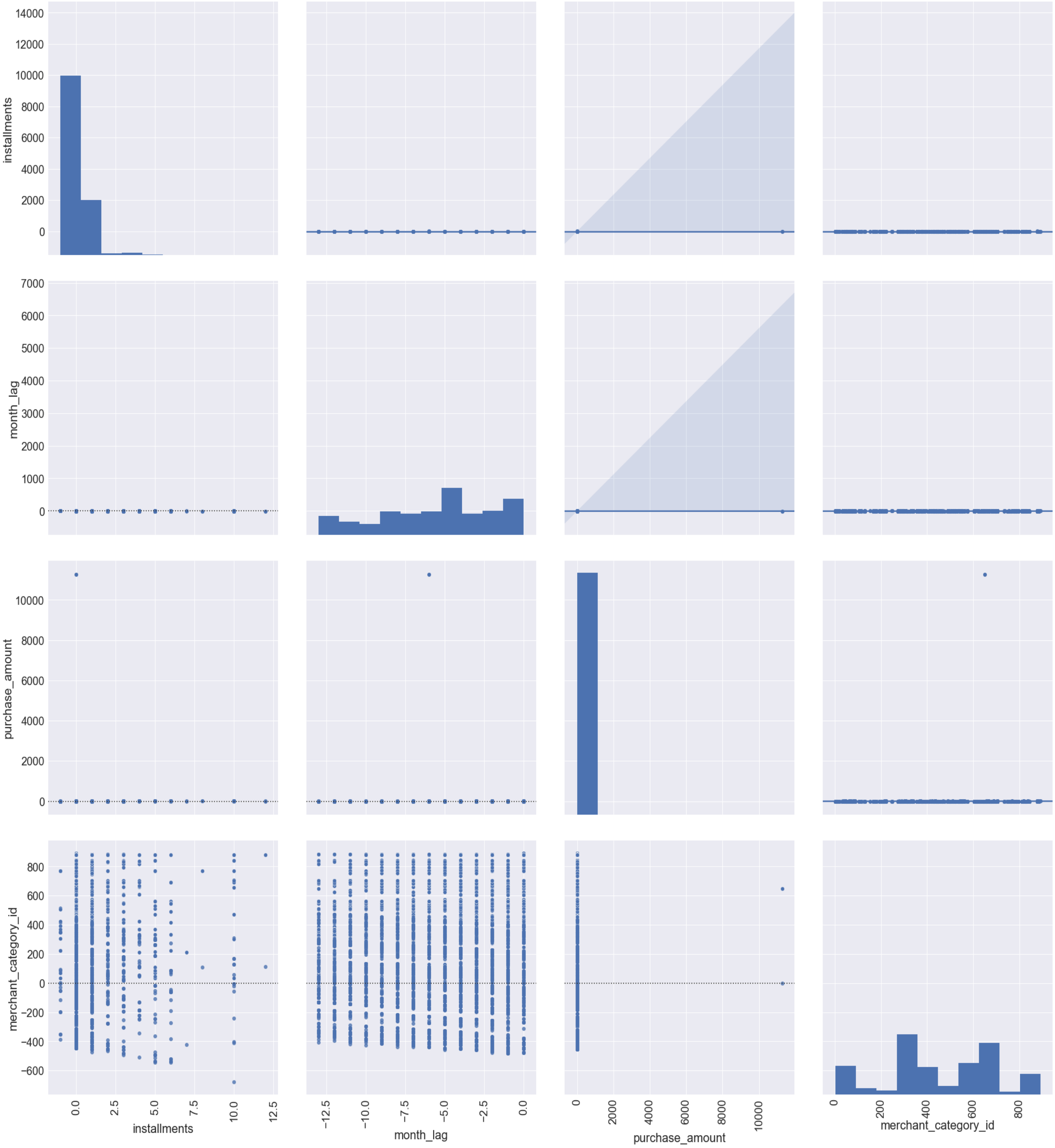
Model Evaluation and Justification:

algorithm	RSME – training set	RSME – testing set
• Linear regression algorithm	3.62	3.61
• Random Forest algorithm	3.86	3.81
• Gradient Boosting Regressor algorithm	3.86	3.86

As can be seen above the linear regression algorithm performs better than others in terms of RSME especially it gets the lower RSME.

Conclusion:

Free-form visualizations:



In the above graph, it shows the distribution of the data for the most features in the data sets.

Reflection:

During our journey in the project , we have done different steps towards Data Science project including the below steps:

Step 0: Import the data and the necessary libraries.
Step 1: Data Exploration and visualization.
step 2: Data cleansing and preparation.
Step 3: Model selection and evaluate the algorithm.
Step 4: Model tuning.
step 5: Results
Further, we have implemented a Data Science model that predict customers' loyalty score. This project is not like other project that we have done during our study in this Nanodegree. We have built a project from a scratch starting by search for a Data Science problem and diagnose it and explore the give data set. Further, date preparation was a little bit tedious and needs to look into different approaches how to prepare the data for the model. Model building and algorithm selection was not easy task, it required a lot data cleansing and preprocessing. Model evaluation was done, and the results was given above.

This project taught that there is still more way to go and learn more , I am planning to participate in more than one Kaggle competition to improve my skills and expose myself to look for a good opportunity.

Improvement:

There are many points and solution could be added to this project to improve:

- I think we could apply neural network algorithms on this project to predict the customers' loyalty score.
- Use time series analysis and regression or apply LSTM model .
- Do more feature engineering especially with date feature and others

Reference:

[1] https://www.kaggle.com/c/elo-merchant-category-recommendation
[2]https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html
[3] https://towardsdatascience.com/linear-regression-using-python-ce21aa90ade6?gi=1e3b5bf86eb5
[4] This project was done by me on Data Science Nanodegree.