Elo Merchant Category Recommendation

My first kaggle competition with big data experience

Competition Timeline

Start Date: November 26, 2018

End Date: February 26, 2019 11:59 PM UTC

Yunfei Bai

UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

Context

- Elo is one of the largest payment brands in Brazil.
- Elo had built machine learning models to understand their customer but none of them is specifically tailored for an individual.
- To develop algorithms to uncover signal in customer loyalty, helping Elo reduce unwanted campaigns and create right experience for customers.

Big data and tools

- Where big data comes in:
  - Data is more than 3GB in total, more than 300K customers for over 30M transactions, can't be read in python as a whole file in 8GB RAM PC.
  - Split into small chunks in order to upload into databrick.

- Tools:
  - Databrick, Scala, Spark Mllib Pipeline, Spark XGBoost

Feature engineering

Merge all historical and new transactions

Filled null data and transform date into numerical variables

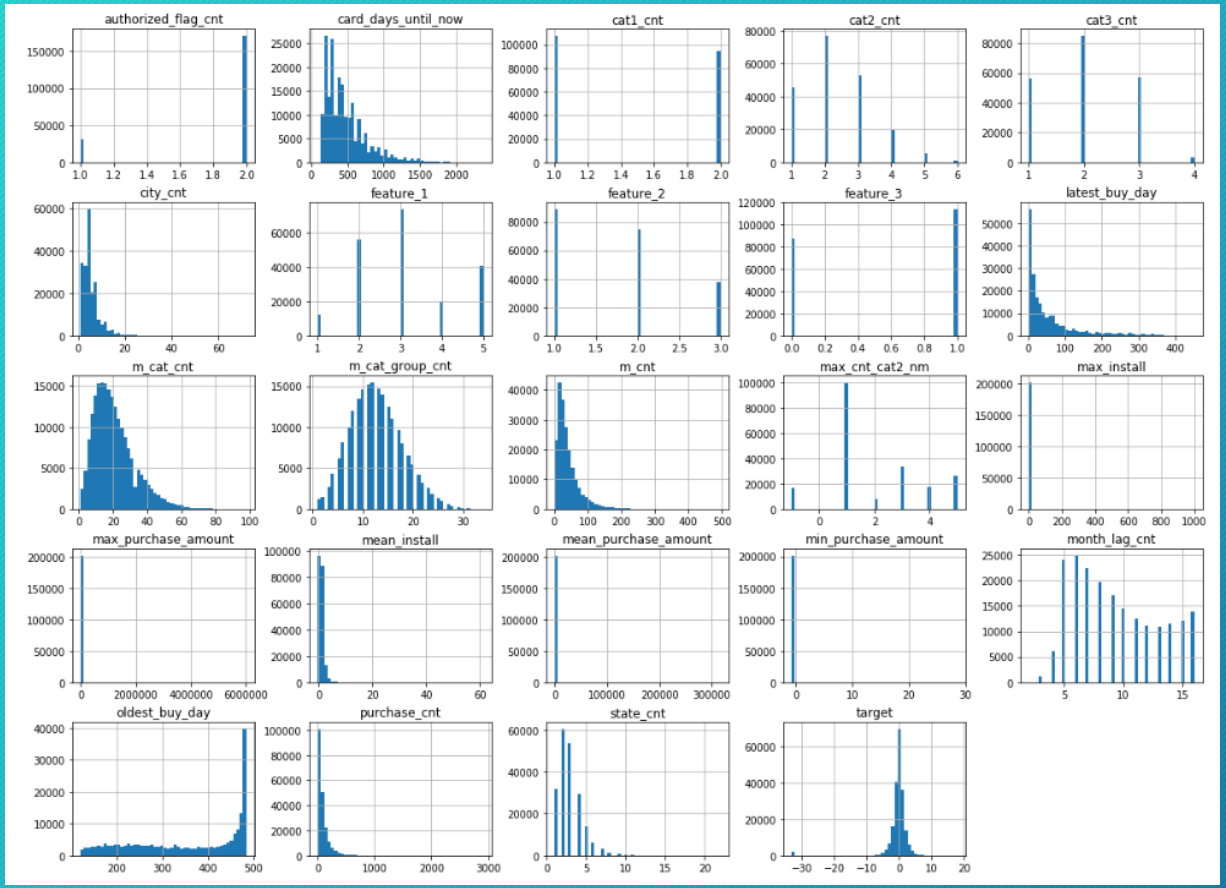
Aggregate transactions into card holder level, and engineer out 19 numerical and 7 categorical features.

Build the pipeline to encoding categorical variables and generate final ensemble feature vector

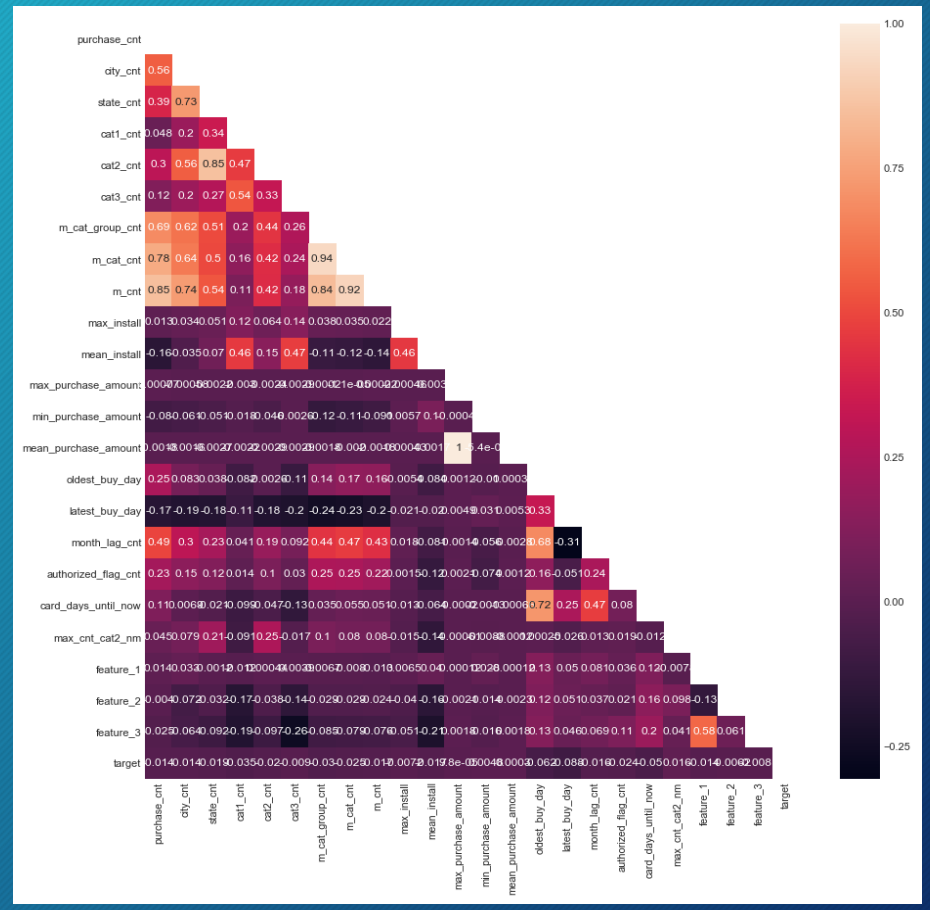
Split the data into 70% to 30% for training and testing

Data Exploration

Numerical features histogram:



Heatmap



Prediction

	Linear Regression	Random Forest Regression	Gradient-boosted Tree Regression	XGBoost Tree Regression
Testing Scores	RMSE : 3.748	RMSE : 3.699	RMSE : 3.702	RMSE : 3.778
LeaderBoard Submission Scores	RMSE : 3.900	RMSE : 3.851	RMSE : 3.838	RMSE : 3.938

A

Radom Forest Regression wins in testing period

B

Gradient Boosting Regression wins in the submission step

My current score and advice

My ranking: around 50%.

560	new	Geekdreams	3.835	6	4d
561	new	chanchino	3.837	2	6d
562	▼188	Tee Ming Yi	3.838	16	3d
563	new	Yunfei Bai	3.838	4	now

Your Best Entry ↑

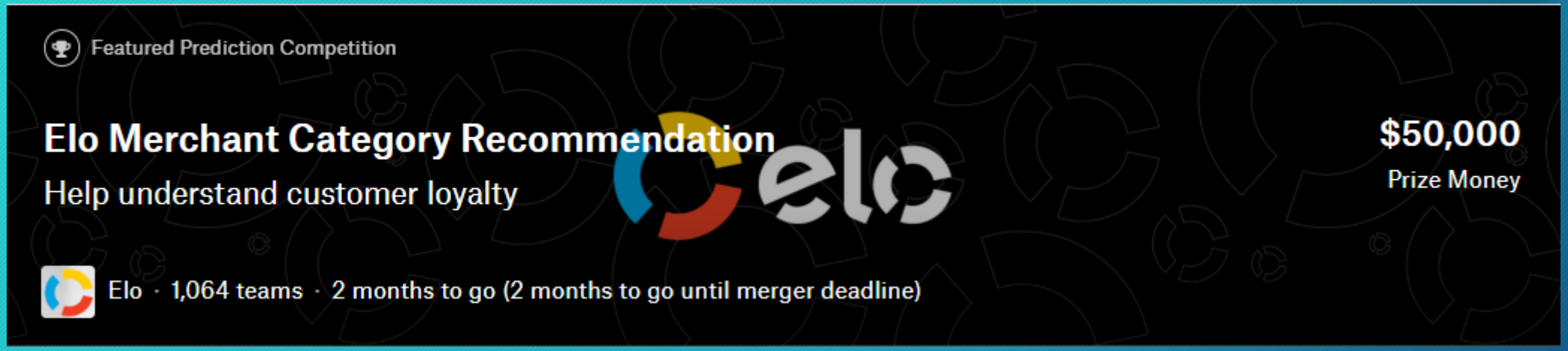
You advanced 10 places on the leaderboard!

Your submission scored 3.838, which is an improvement of your previous score of 3.851. Great job! Tweet this!

564	▼346	Mikhail Novikov	3.838	1	10d
565	▼197	I'll be on the LB one day	3.839	6	1d

Advice: so far the best scores are coming from 5-Fold LightGB Model.

Good luck and have fun!



Databrick notebook:
https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/7792953078525217/2896759813459456/4057496721065160/latest.html
Github link: https://github.com/yunfeibai123/3252-Class-Project

UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES