Machine Learning Engineer Nanodegree
Elo Merchant Category Recommendation

Joël Holla
j_h_olla@hotmail.com

December 10, 2018

Proposal

Domain Background:

[1] Imagine being in a place that you have never been to before and getting restaurant recommendations served up, based on your personal preferences. The recommendation comes with an attached discount from your credit card provider for a local place around the corner. Companies such as Elo, one of the largest payment brands in Brazil, has built partnerships with merchants in order to offer promotions or discounts to cardholders. We must ask ourselves the questions do these promotions work for either the consumer or the merchant? Does the consumer enjoy their experience? Personalization is the key that will help us unlock these questions. Elo has built machine learning models to understand the most important aspects and preferences in their customers' lifecycle. So far none of these models is specifically tailored for an individual or profile.

[2] Personalized marketing, also known as one-to-one marketing is one of the many tools that is being used by companies. Individual marketing is a marketing strategy by which companies leverage data analysis and digital technology to deliver personalized messages and products offerings. This strategy is dependent on many different types of technologies for data collection, data classification, data analysis, data transfer, and data scalability.

This type of marketing is not only a benefit for both the different businesses and companies, it also can greatly aide and save time for the consumer. Before the Internet, consumers did not face such a wide range of variety and volume of products and services. In this modern age, however for example, a single retail website can offer thousands of different products and services. Personalized marketing helps to bridge the gap between the vastness of what is available and the needs of individuals consumers, which can greatly enhance the consumer's experience.

A few examples of machine learning problems that are similar to this project are the following:
[3] IBM, through the use of IBM Watson and the plethora of customer data, use the power of artificial intelligence (e.g. predictive analytics) to improve how they manage customer relationships to increase customer loyalty. This is done by examining the different type of loyalty behaviours and the factors associated with it.
[4] Airbnb, a leader among p2p residential real estate focus their efforts on the enhancing the search experience, by using the AI models to analyze more than a hundred signals at once, to help personalize the search in real time.
[5] A study that was done in 2007, by Wouter Buckinx, Geert Verstraeten and Dirk Van den Poel, predicting customer loyalty using the internal transactional database.
[6] A project by Ashish Gandhe describes the work to learn to predict whether a given yelp user visiting a restaurant will like it or not.

My personal motivation with this project is to see how data can aid businesses sectors, such as retail, grow and adapt to this ever-changing world.

Problem Statement:

In this project, I will develop an algorithm to identify and serve the most relevant opportunities to individuals, by uncovering signal in customer loyalty. I will be doing this by predicting a *loyalty score* for each card_id represented in the **test.csv.** This will give a business such as Elo a way to better keep track of customer loyalty which will allow them to reduce the number of unwanted campaigns and to create the right experience for their customers.

Dataset and Inputs:

The dataset contains 5 .csv filesl, you will need at a minimum the **train.csv** and **test.csv** files. These contain the card_id's that I will using for training and predictions. There is also **historical_transactions.csv, new_merchant_transactions.csv** and **merchants.csv.**

- The **historical_transactions.csv** contains up to 3 months' worth of transactions for every card at any of the provided merchant_ids.
- The **new_merchant_transaction.csv** contains the transactions at *new* merchants (merchant_ids that this particular card_id has not yet visited) over a period of two months.
- The **merchants.csv** contains aggregate information for each merchant_id representd in the data set.

1. **train.csv** and **test.csv** - contains

    - card_id – the id of the different credit cards
    - first_active_month – list of transactions made by the different card_ids
    - feature_1 to feature_3 – discrete variables (the range of values ranges from 0 to 5, 0 to 3, and 0 to 1 respectively)
    - target – the loyalty score for that particular card_id. This is the target variable. In test.csv, target is not present since we are going to predict this variable.

2. In **trains.csv:**
    - Number of rows = 201917
    - Number of columns = 6
    - Highly relevant as this is the data we will train on.

3. In **test.csv**:
    - Number of rows = 123623
    - Number of columns = 5
    - Highly relevant as this is the data, we will test our model on.

Historical Transactions

card_id	Card identifier
month_lag	month lag to reference date
purchase_date	Purchase date
authorized_flag	Y' if approved, 'N' if denied

category_3	anonymized category
installments	number of installments of purchase
category_1	anonymized category
merchant_category_id	Merchant category identifier (anonymized)
subsector_id	Merchant category group identifier (anonymized)
merchant_id	Merchant identifier (anonymized)
purchase_amount	Normalized purchase amount
city_id	City identifier (anonymized)

New merchants transactions

state_id	State identifier (anonymized)
category_2	anonymized category

New merchants transactions

card_id	Card identifier
month_lag	month lag to reference date
purchase_date	Purchase date
authorized_flag	Y' if approved, 'N' if denied
category_3	anonymized category
installments	number of installments of purchase
category_1	anonymized category
merchant_category_id	Merchant category identifier (anonymized)
subsector_id	Merchant category group identifier (anonymized)
merchant_id	Merchant identifier (anonymized)
purchase_amount	Normalized purchase amount
city_id	City identifier (anonymized)
state_id	State identifier (anonymized)
category_2	anonymized category

As this was a Kaggle competition. The dataset is provided by Kaggle and Elo. They can be obtained here.

Solution statement:

The preprocessing done in “Prepare data” section of the notebook consist of the following steps:

1. Check the quality of the data given and perform data cleaning.
2. We prepare the data by splitting the features and target columns.
3. Splitting the datasets into a training set and a validation set. This was already done by Kaggle which provided the dataset.

Verifying the quality of the data is good, because if there are any missing values, that could great affect the outcomes of the model during training. Several things that are done during data cleaning is also checking to see if there are any non-numeric columns that need to be converted, machine learning models learn on numeric values. Columns that have more than two values are known as categorical variables, the recommended way to handle such columns is to create as many columns as needed of possible values and assign a 1 to one of the values and 0 to all others possible values. These generate columns that are sometimes called dummy variables, and which I will use the pandas.get_dummies() function to perform this transformation.

This problem is a regression with a larger dataset with a couple hundred thousand rows. I don’t know which algorithm would be a good fit for this problem or what configuration would be to use. Here are a few algorithms to evaluate.

1. Ridge Regression
2. LightGBM
3. XGBoost(XGB)

I will also be using 5 folds cross validation to estimate root mean squared error. This will split our dataset into 5 parts, the data will train on 4 parts and test on 1 part. This will repeat through all combinations of train-test split.

We will be using the metric of root mean squared error to evaluate the models. This is to see the amount of error between the predicted value and the actual values.

Benchmark:

I will be using a simple linear regression as a benchmark, to use as a baseline score for my dataset when comparing my how well my finalized model performed. Also since this is a Kaggle competition another good benchmark would be the best Kaggle score for the test set, which comes 3.713 root mean squared error (RMSE) lower is better.

The mathematical definition for RMSE can be defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

where,
n = number of observations
y_i = actual value of target variable
$y_i\ hat$ = predicted values of target variable

In this project, I will calculate the RMSE, by calculating the square root of mean square error () function provided in the metrics module of sci-kit learn library.

Evaluation Metrics:

The model prediction for this problem can be evaluated in several ways. Since the official evaluation of this project is done by Kaggle using the root mean squared error (lower it is, the better the model), same evaluation metric will be used in this project.

Project design:

The workflow for solving this problem will be in the following order:

- Exploring the data
  - Loading the necessary libraries
  - Examine the different datasets
  - Exploring the dimension of the data
  - Statistical summary
- Data preprocessing/cleaning
  - Data cleaning
  - Identify feature and Target columns
  - There is no need to split the data that part was already performed by Kaggle
  - Merge the different datasets together
  - Feature scaling – Standardization/normalizing data
- Evaluate Algorithms
  - Build models
  - Select best model
  - Make predictions on the validation set
  - Determine feature importance
  - Do feature selection
- Model Tuning to improve result
- Final conclusion

References:

[1] https://www.kaggle.com/c/elo-merchant-category-recommendation

[2] https://en.wikipedia.org/wiki/Personalized_marketing

[3] http://businessoverbroadway.com/2018/03/19/using-predictive-analytics-and-artificial-intelligence-to-improve-customer-loyalty/

[4] https://techspective.net/2018/05/16/how-leading-companies-use-ai-for-customer-retention/

[5] https://www.google.ca/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=2ahUKEwilmrSotZnfAhVM5YMKHdhvCKoQFjABegQIBxAC&url=https%3A%2F%2Fwww.u-cursos.cl%2Fingenieria%2F2010%2F2%2FIN71K%2F1%2Fmaterial_docente%2Fbajar%3Fid_material%3D307341&usg=AOvVaw2hkrOsW5KIRl0Ob6mVkn60

[6] http://cs229.stanford.edu/proj2014/Ashish%20Gandhe,Restaurant%20Recommendation%20System.pdf