

Churn Prediction

Find Risky account and build model for churn prediction

**Prajakta Gujarathi | Marqeta-Data analyst/scientist Application | 3/25/2018**

Assumptions & Problem Statement

## Problem Statement

The focus on customer churn is to determinate the customers who are at risk of leaving and if possible on the analysis whether those customers are worth retaining. The churn analysis is highly dependent on the definition of the customer churn. The business sector and customer relationship affects the outcome how churning customers are detected. Example in credit card business customers can easily start using another credit card, so the only indicator for the previous card company is declining transactions.

So project goal is, given a data of spend and count per account we are creating supervised learning binary classification model which will  predict if given account is at risk of churn of not.

## Churning Definition:

For this project I am defining churning by three factors if there is not activity for last 4 months and if amount of transaction and count of truncations are decreasing over the period then those account are at the risk of churning.

## Datasets and Inputs:

• Spend ( [https://www.dropbox.com/s/qia89hd3i2j2nl6/spend.csv.gz?dl=0](https://www.google.com/url?q=https://www.dropbox.com/s/qia89hd3i2j2nl6/spend.csv.gz?dl%3D0&sa=D&source=hangouts&ust=1522119066568000&usg=AFQjCNE-kADsgl6UViwJ5kPAEkFySIraPQ) )  
• Counts ( [https://www.dropbox.com/s/jbnzgk6shgnz5hb/counts.csv.gz?dl=0](https://www.google.com/url?q=https://www.dropbox.com/s/jbnzgk6shgnz5hb/counts.csv.gz?dl%3D0&sa=D&source=hangouts&ust=1522119066568000&usg=AFQjCNEg1QRu6tAGYDy3ldplBXGVowbc5A) )  
The variables between the two files are:  
• date   
• account id   
• amount of spend for a time period   
• number of transactions for a time period

Analysis and Methodology

To start with this analysis we will need to explore, clean and manipulate the data.

Luckily we don’t have missing data.

**There are 10,000 unique accounts.**

Amount Feature Analysis:

amount

count 1.934M

mean 301.837

std 1.122k

min -605.167

25% 15.327

50% 66.135

75% 234.994

max 307.370k

Here we can see the difference between mean and median is quite large also the max and min values seems to be outliers. **Not normally distributed median is not affected by outliers**, so median is best indicative of spending for amount feature.

Count Feature Analysis:

count

count 1.934M

mean 41.459

std 74.995

min 10.000

25% 15.000

50% 24.000

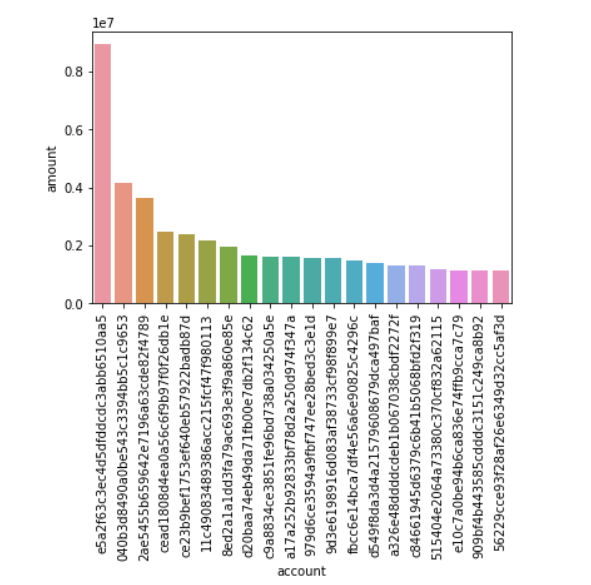
75% 44.000

max 5.643k

Count seems to be normally distributed with some outliers in it like max value seems to be abnormal.

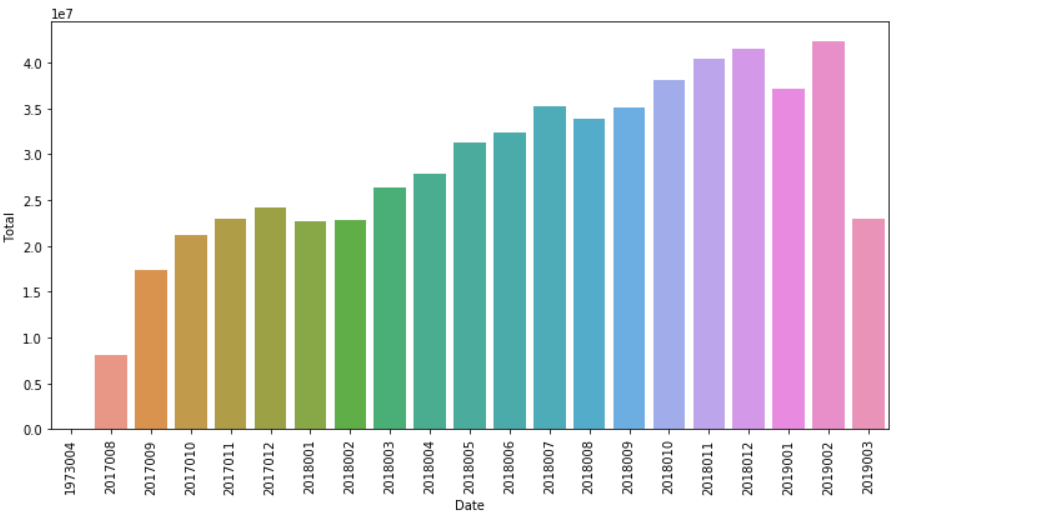
Most Spending by account:

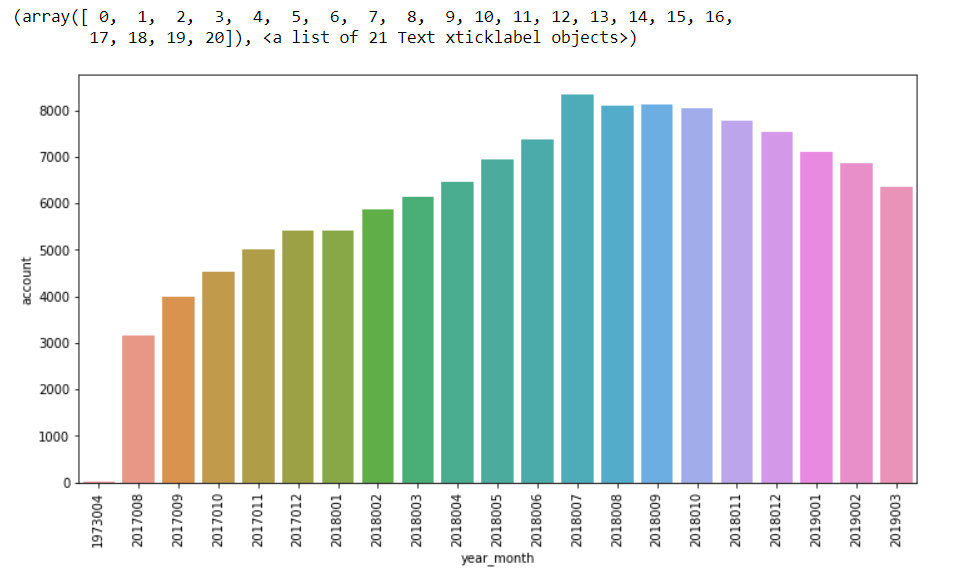
Here are top 20 accounts who has max spending over the given period with average 25k per account. (As spend amount data is **not normally distributed** I am considering **median** as my average amount)



Total Monthly Spending:

Graph below shows that total monthly spending are increase with time.



Number of active account per Month: Below graph shows that 2018/7 has most active account, till them number of account are increasing and count drops after that.

Similarly data exploration is done on count data set.

# Feature creation and feature Transformation:

In this step I merged data from both the data set on date and account columns.

Amount related features:

* Average amount per account
* Quarterly average amount spend per year

Date related features:

* Start date
* End Date
* Total active days in a period (this can be most important feature in predicting churn)
* Days seen Last transaction
* Quarterly average active days per year

Count related features:

* Average transaction per account
* Quarterly average transaction count per year

Churn related Features:

* Risky by last activity(if days seen last transaction is greater than 4 month then value is 1 else 0, found 2074 such accounts)
* Risky by decrease Average transaction count (so if there is gradual decrease in transaction count per quarter. Found 41 such accounts)
* Risky by decrease Average spending amount (so if there is gradual decrease in spend amount per quarter. Found 35 such accounts)
* Total weight (it is calculated by giving weightage to Risky by last activity 50%, Risky by decrease Average transaction count 25% and Risky by decrease Average spending amount 25%)

**Risky account can be found in two ways:**

1. **Equally weighing each churn features:** So if either of the churn feature (Risky be last activity, risky by decrease average transaction and risky bye decrease average spending**)** is 1 then account is labelled as risky.
2. **Unequal weightage for each churn features:** I am going to use weights for each feature, days since last activity 50%, decreasing trend per quarter for amount 25% and decreasing trend per quarter for number of transaction 25%.

**List of risky account can be found in file risky\_account.csv**

Model Building

I consider this as classification problem where 0 classifies as non-risky accounts 1 classifies as risky ones.

# Metrics:

The data we have is going to be unbalanced data i.e. number of non-risky accounts (class 0) are always going to be more than those risky (class 1).

### Evaluation Metrics:

F1 is a common metric for binary classifiers and for unbalanced data.

PRECISION = TP / TP+FP

RECALL = TP/ TP + FN

F1 = 2(PRECISION \* RECALL) / (PRECISION + RECALL)

TP = Account which are actually risky

FP = Account which are predicted to be risky but are not

TN = Account not risky

FN = Account are risky but predicted as not risky.

F1 score is optimal choice for this problem as F1 actually doesn't weight TP and TN equally. It disregards true negatives because they're likely to be far more common like since we could simply say "Account not risky" and generally be right, without ever looking at the data, and thus not as critical to get right.

### Model Creation:

In this step, I apply various supervised learning models on preprocessed data like Logistic regression, Knn, RandomForestClassifer etc. by dividing training data set into train and test.

Comparing the F1 score for each model and chose the one with the highest F1 score. Also, considerd model performance - how much time each model took to train and learn.



Chose the best model as RandomforestClassifer

This model is saved in file **RandomForest\_equally\_weighted\_model.sav and RandomForest\_unequally\_weighted\_model.sav**

Once the model is finalized using grid search, tuned the model parameters.

Selected the best parameters and use that model to predict the test data.

Below are top 5 important features which helps decide which accounts are in risk of churning.¶

1. 2018004\_avg\_tran (0.208727)

2. total\_weights (0.197219)

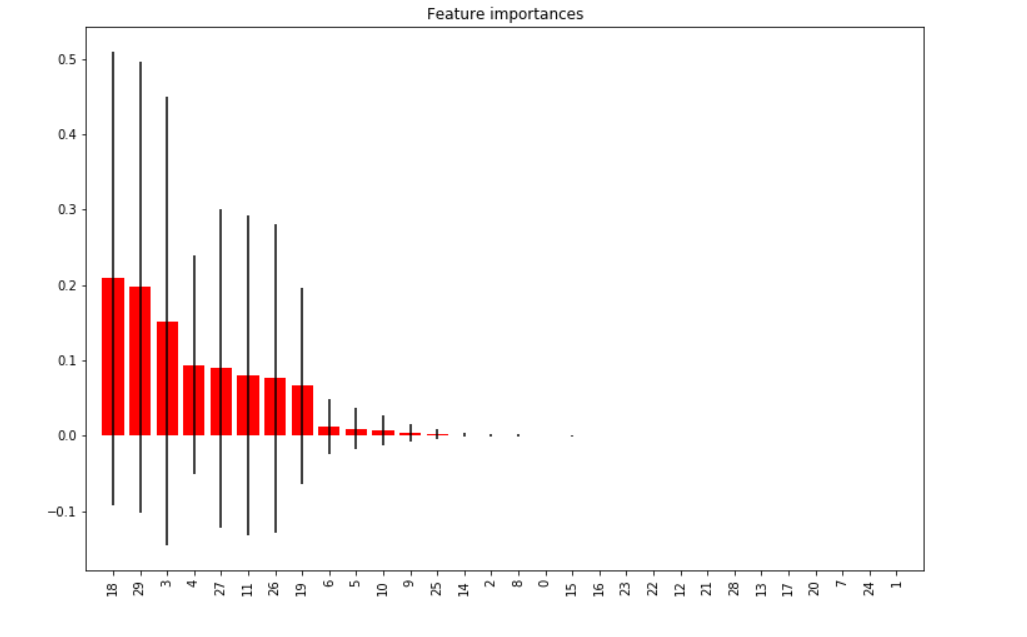
3. days\_seen\_last\_transaction (0.151545)

4. risky\_by\_last\_activity (0.093938)

5. 2019001\_days (0.089488)

We have found important features which can point out risky account. But the data is not sufficient to make predications. We need more data to rely on model prediction like demographic information, type of transactions, age of credit card, number of other credit cards etc

**Feature Importance:**



Now this model can be used to predict result for unseen data.

Expected Results and Findings

So given the inputs as account id and transaction amount per quarter and count to this model, it is expected that this model will predict the class for that account, i.e. if the account is risky or not.

Depending upon the business need and churn definition we can build a model to predict churning accounts.