# OFFICE SUPPLY STORE DATA ANALYSIS

June, 2020

By: Juan Sebastian Pinzon Gamez

# AGENDA

- Background
- Objectives
- Methodology
- Exploratory Data Analysis
- Classifier model
- Regressor model
- Profitability assessment
- Decile profile
- Recommendations
- Appendix

# BACKGROUND

- An office supply store is planning to test a telemarketing campaign to its existing business customers.
- The company has generated real data with approximately 16,000 customers for the campaign.
- The client wants to use the response data from a previous campaign to improve the response of the next campaign and maximize its profit.
- The products being marketed are Desk, Executive Chair, Standard Chair, Monitor, Printer Computer, Insurance, Toner and Office Supplies .

- Database variable:
  - Prior campaign sales
  - Historical sales
  - Prior year transactions
  - Date of first purchase
  - Targeted customers
    - Number of employees
    - Language
    - Products purchased
    - Transaction channels

# OBJECTIVES

- Develop models that will allow the company to use the results of the campaign to target responsive, profitable customers for future campaigns.
- Provide financial value of the models based on profitability which may be used to optimize future campaigns.
- Profile the customers that are probable responders to the campaign to understand the characteristics of them.

# METHODOLOGY

1. Perform Exploratory Data Analysis on the dataset obtained from the company.

2. Transform the dataset:
   - Create training and testing datasets (50/50).
   - Imputing.
   - Standardization.
   - Feature engineering.

3. Creation and validation of data models
   - Random Forest Classifier model to estimate the probability of responding to this campaign.
   - Gradient Boosting Regressor model to determine expected amount ($) of the transaction.
   - Calculate expected profitability from models results.

4. Define customer deciles.

5. Create lift table.

6. Deliver recommendations

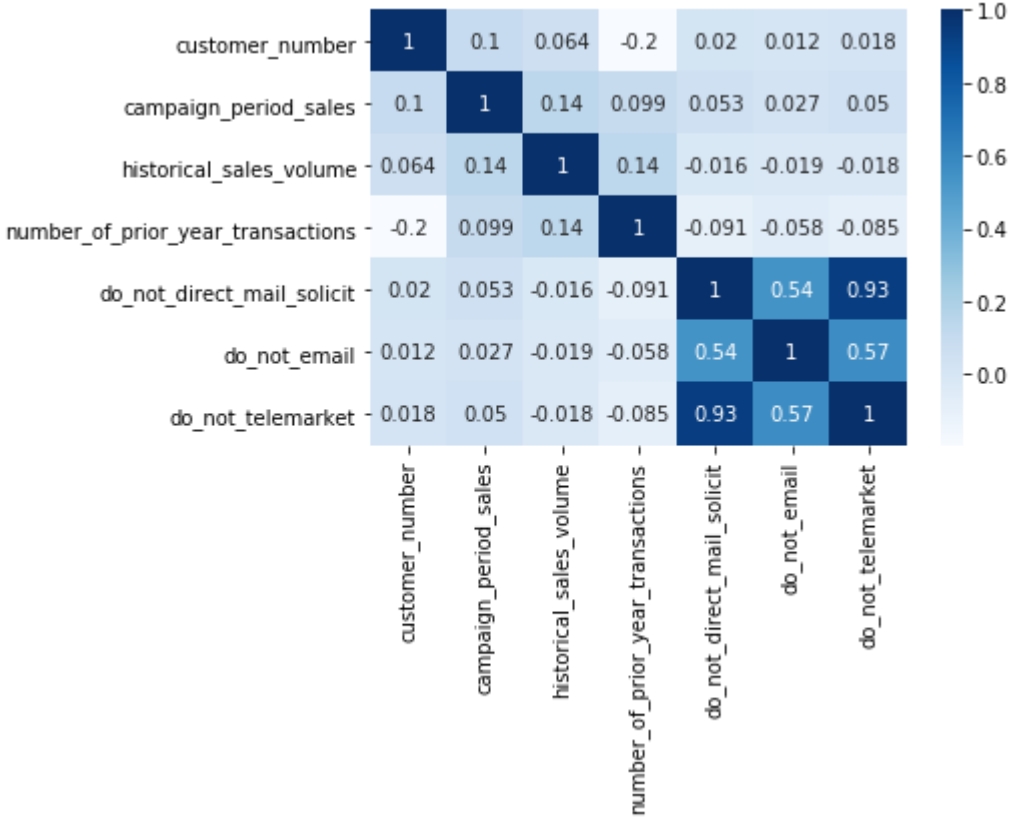
# EXPLORATORY DATA ANALYSIS

- Shape of data
  - 16173 customers
  - 21 features
- Check for null values for customers
  - 1 dropped
- Check for duplicate customer's ID.
- Check for correlations
  - Variables have low correlation
- Negatives values
  - Campaign sales (6)
  - Historical sales (4)

- Check for information:

| | Mean | Max |
|---|---|---|
| Campaign sales ($) | 245.89 | 8936.85 |
| Historic sales ($) | 671676.3 | 34.41 M |
| Prior year transactions | 14.48 | 313 |

- Findings from last campaign:
  - Customers that purchased = 4379
  - Total campaign sales: 3.97 millions
  - Oldest customer that purchased: since 1946
  - 55 customers purchased all products ($3190 on avg)



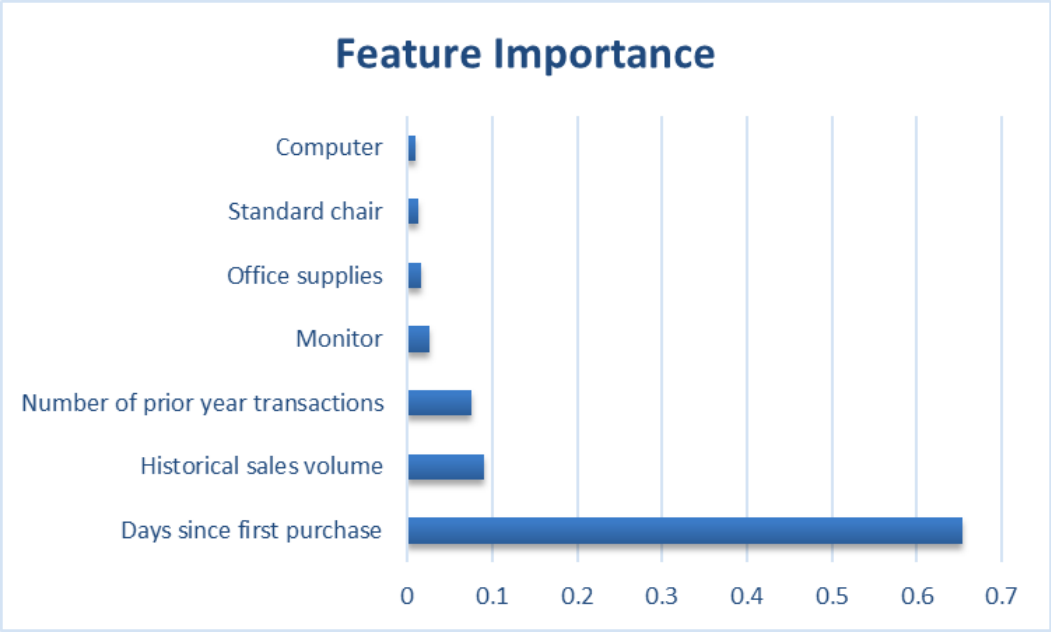
# CLASSIFIER MODEL

Model:

- Random Forest Classifier to determine contribution or not.

Setup:

- Model built using 50/50 split on existing campaign customers.
- Validation Sample Size: 8086




- The model accurately predicts the purchase 87% of the time when responding to the campaign.
- Numerical features are the most relevant for the model where customer's seniority is the most important feature to target responsive and profitable customers for the upcoming campaign.

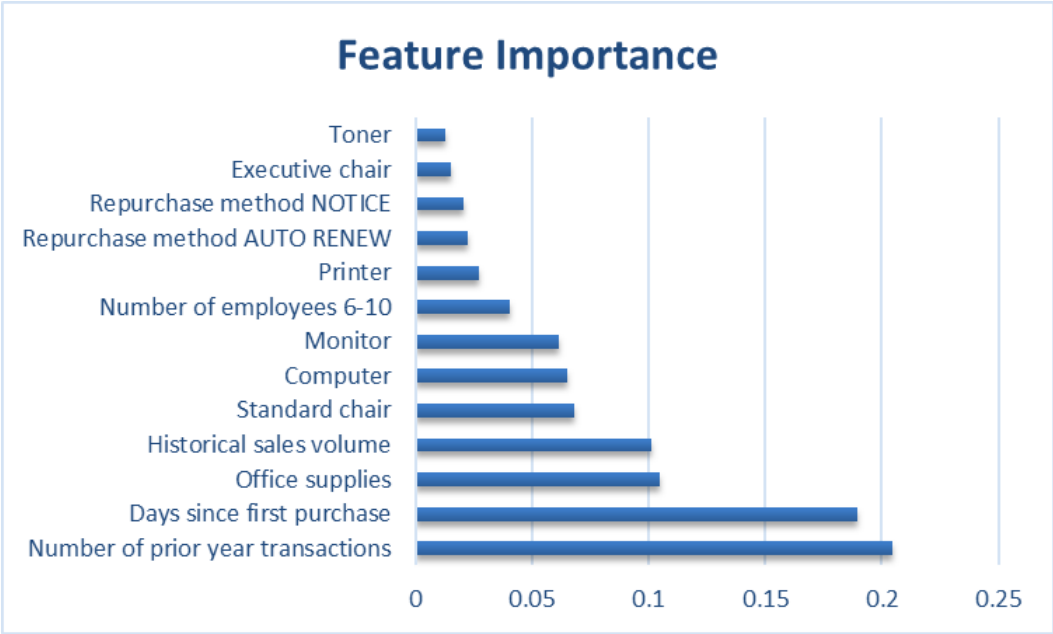
# REGRESSOR MODEL

Model:

- Gradient Boosting Regressor to determine amount of the contribution ($).

Setup:

- Model built using 50/50 split on existing campaign customers.
- Validation Sample Size: 8086




- In this case, several features are relevant to determine the expected amount ($) of the transaction, being the prior year transactions and the customer's seniority the most important ones (~40% of contribution).

- Office supplies and Historical Sales contribute with ~20%.

- Due to the shape of the data (having ~73% of 0's), when validating the model, we are over forecasting small values and under forecasting larger values. This is simply a shortcoming of the model since we are usually trying to predict around the expected value of our data.

# PROFITABILITY ASSESSMENT

Expected campaign profitability was determined utilizing the following estimated financial metrics:

- Gross margin on sales: 22%.
- The campaign cost: $45.65 per customer contacted by the sales force.
- The transaction cost: $8.40 per transaction.

$$E(Profit)=.22*Prob(Sale)*Est(Transaction\ Size)-\$8.40*Prob(Sale)-\$45.65$$

Probability of sale obtained from the classifier model and the estimated transaction size obtained from the regressor model were combined with the cost and margin data to assign a predicted profit to all customers in the validation set.

The customer base result (8000) was scored and ranked into 10 deciles (800 each) to perform a lyft analysis to measure how the campaign impacts costs and profit for the company.

# PROFITABILITY ASSESSMENT

| Decile | Number of Customers | Actual Profitabilty Per Customer | Lift Over Average | Total Profit | % of Profit | Incr Proj Profit 100k Cust Base ($K) | Total Proj Profit 100k Cust Base ($K) | Cuml Incr Profit 100k Cust Base ($K) | Cuml Total Profit 100k Cust Base ($K) |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 800 | $ 189 | $ 204 | $ 151,312 | 946% | $ 2,039 | $ 1,891 | $ 2,039 | $ 1,891 |
| 2 | 800 | $ (6) | $ 9 | $ (4,512) | -28% | $ 92 | $ (56) | $ 2,131 | $ 1,835 |
| 3 | 800 | $ (25) | $ (10) | $ (20,016) | -125% | $ (102) | $ (250) | $ 2,029 | $ 1,585 |
| 4 | 800 | $ (39) | $ (24) | $ (30,936) | -193% | $ (239) | $ (387) | $ 1,790 | $ 1,198 |
| 5 | 800 | $ (42) | $ (27) | $ (33,760) | -211% | $ (274) | $ (422) | $ 1,516 | $ 776 |
| 6 | 800 | $ (44) | $ (29) | $ (34,976) | -219% | $ (289) | $ (437) | $ 1,227 | $ 339 |
| 7 | 800 | $ (45) | $ (30) | $ (36,032) | -225% | $ (302) | $ (450) | $ 925 | $ (112) |
| 8 | 800 | $ (46) | $ (31) | $ (36,408) | -228% | $ (307) | $ (455) | $ 617 | $ (567) |
| 9 | 800 | $ (46) | $ (31) | $ (36,520) | -228% | $ (309) | $ (457) | $ 309 | $ (1,023) |
| 10 | 800 | $ (46) | $ (31) | $ (36,624) | -229% | $ (310) | $ (458) | $ (1) | $ (1,481) |
| Total | 8k | $ (14.8) | $ (0) | $ (118,472) | -740% | $ (0) | | $ (1) | |

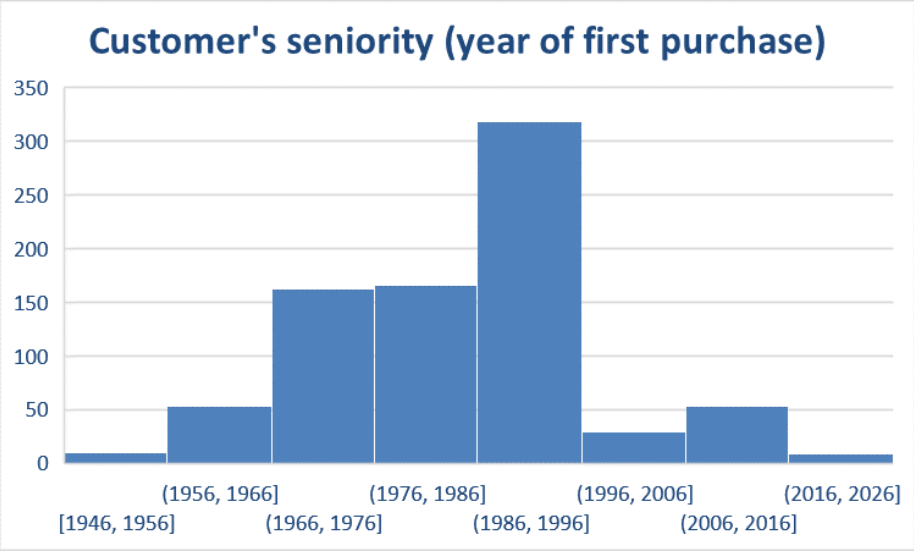
If targeting all 8000 customer, the marketing campaign would generate an overall loss per customer of $14.8.

The first decile is profitable with an average profitability of $189 per customer and a total profit of $151312.

Targeting the marketing campaign to the first decile will generate a forecasted profit of $2.039 million on a 100K customer base.

The second decile has a potential increment in project profit of $92K. From the 800 customers in the second decile, there are 201 customers with positive expected profit.

# DECILE PROFILE




| Number of Employees | Companies |
|---|---|
| 1 - 5 | 12 |
| 6 - 10 | 122 |
| 11 - 50 | 44 |
| 51 - 100 | 47 |
| 101 - 500 | 49 |
| 500+ | 29 |

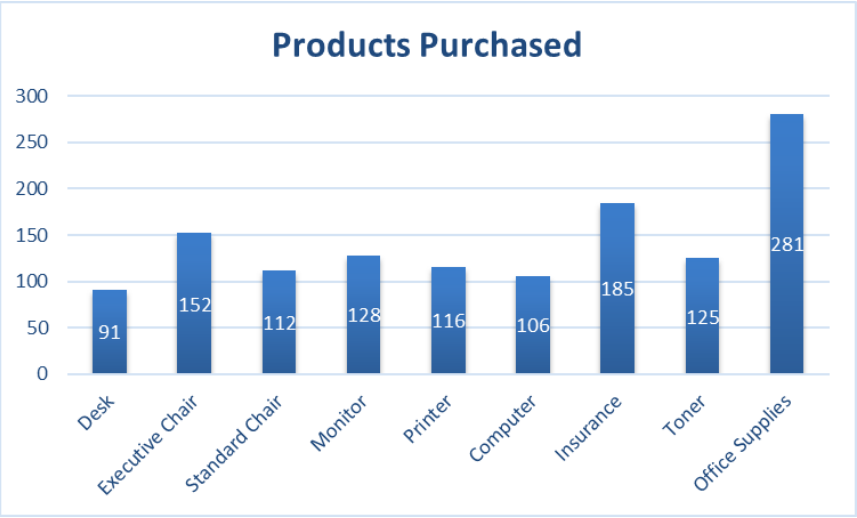
| Language | Companies |
|---|---|
| English | 261 |
| Spanish | 3 |
| Portuguese | 1 |
| Polish | 1 |

318 customers from the decile made the first purchase between 1986 and 1996, having a seniority of around 30 years.

From the 800 customers on the decile, you have information about Number of Employees from 303 companies, where 122 companies have between 6 and 10 employees and 29 are large companies with more than 500 employees.

From the information provided, in the first decile there are 261 companies which their language is English.

# DECILE PROFILE




From the 800 customers, 612 purchased products on the last campaign, with sales around $1.2 millions.

Office supplies were the most purchased item followed by insurance.

| | Average of values by size of company | | | |
|---|---|---|---|---|
| Number of Employees | Campaign Period Sales | Historical Sales Volume | Prior Year Transactions | Years since 1st Purchase |
| **1-5** | $162.51 | $313,271.65 | 17.00 | 45.63 |
| **6-10** | $251.58 | $709,783.76 | 19.40 | 36.44 |
| **11-50** | $490.14 | $1,044,477.48 | 18.66 | 33.03 |
| **51-100** | $577.15 | $1,293,063.29 | 18.00 | 36.60 |
| **101-500** | $786.23 | $2,001,914.18 | 17.14 | 34.56 |
| **500+** | $1,562.77 | $3,272,262.82 | 17.55 | 32.63 |

Larger companies from the decile have purchased more historically and in the last campaign. Smaller companies with less than 1-5 have more seniority than the others. In general, all companies made a similar amount of transactions in the prior year.

# RECOMMENDATIONS

- Customers on the first decile must be targeted by the sales force in the next campaign to maximize profitability (potential profit $2M).

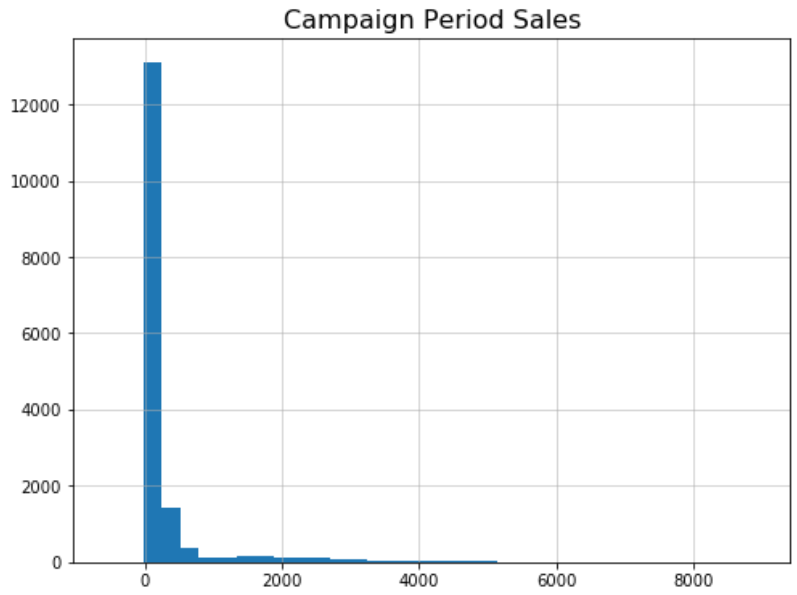
- Attempting to target the second decile could generate a loss in profit, but on a 100k customer base it has a potential increment project profit of $92K. The recommendation is to make a selection on the possible profitable customers in this decile or test lower cost communication methods such as web or email.

- Update the missing information on the database from customers, such as Number of Employees and Language, and try to obtain the complete information from new ones.

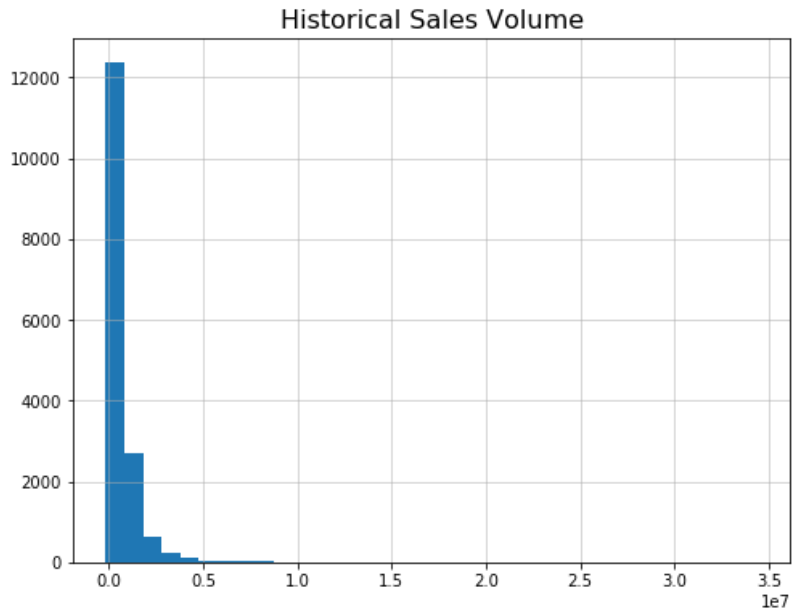
- To maintain and improve the models, an analyst could be hired, and he/she can perform some consolidation on the variables and advanced feature engineering to refine the database and models.

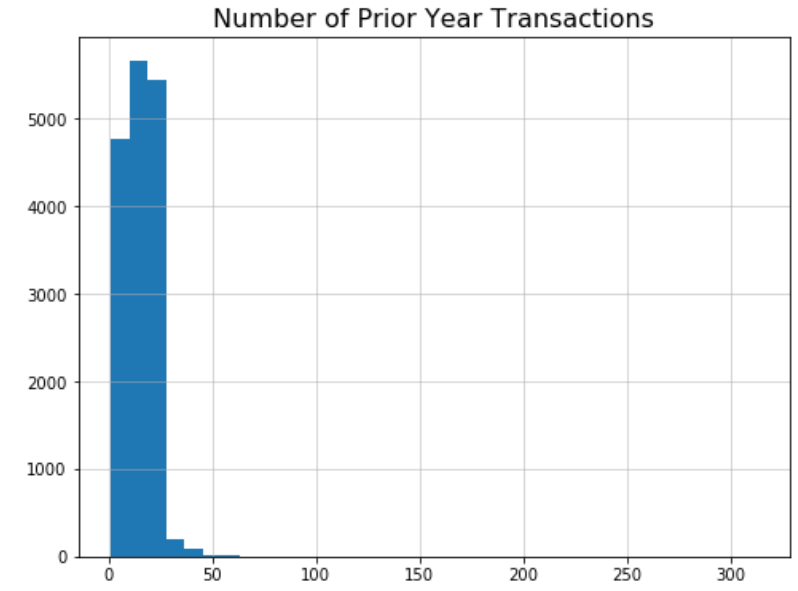
# APPENDIX

- Distribution of target variable




- Distribution of numeric variables



Classifier Model:

Random Forest Classifier

Model evaluation:

Score train set: 0.9063813999505318

Score test set: 0.8660648033638387

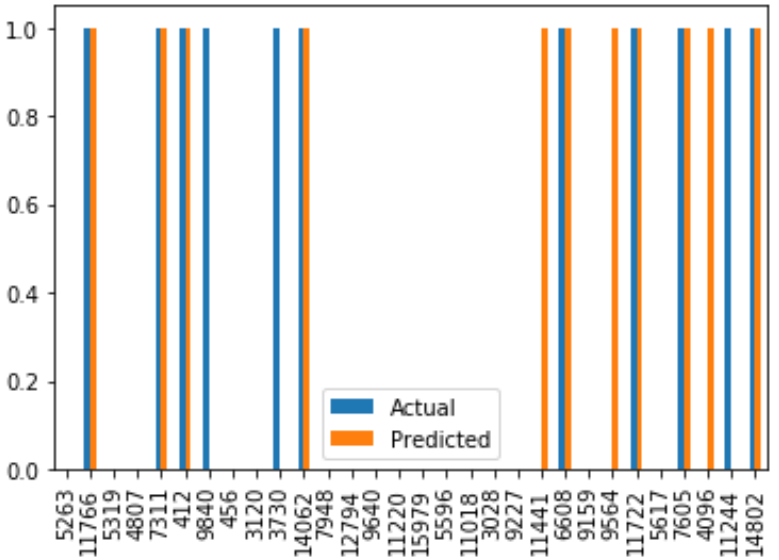
Cross val score: [0.868356 0.86641929 0.85776129 0.85899814 0.86641929]

Cross val precision: [0.77506112 0.80108992 0.78201635 0.75802469 0.80662983]

Class Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.92 | 0.96 | 0.94 | 5899 |
| 1 | 0.87 | 0.77 | 0.82 | 2187 |

- Evaluation of sample of predictions



Regressor Model:
GradientBoostingRegressor

Model evaluation:
RMSE train: 359.21837507312966
R2 train: 0.7631641261008051

Cross val RMSE(mean) train: 504.2603767901928

RMSE test: 464.9770110886481
R2 test: 0.5610552531461248

- Evaluation of predictions