Emine Akbulut,

akbulutemine7@gmail.com

757-500-1667

CREDIT CARD ANALYSIS OF PSPD BANK

MARCH 15, 2020

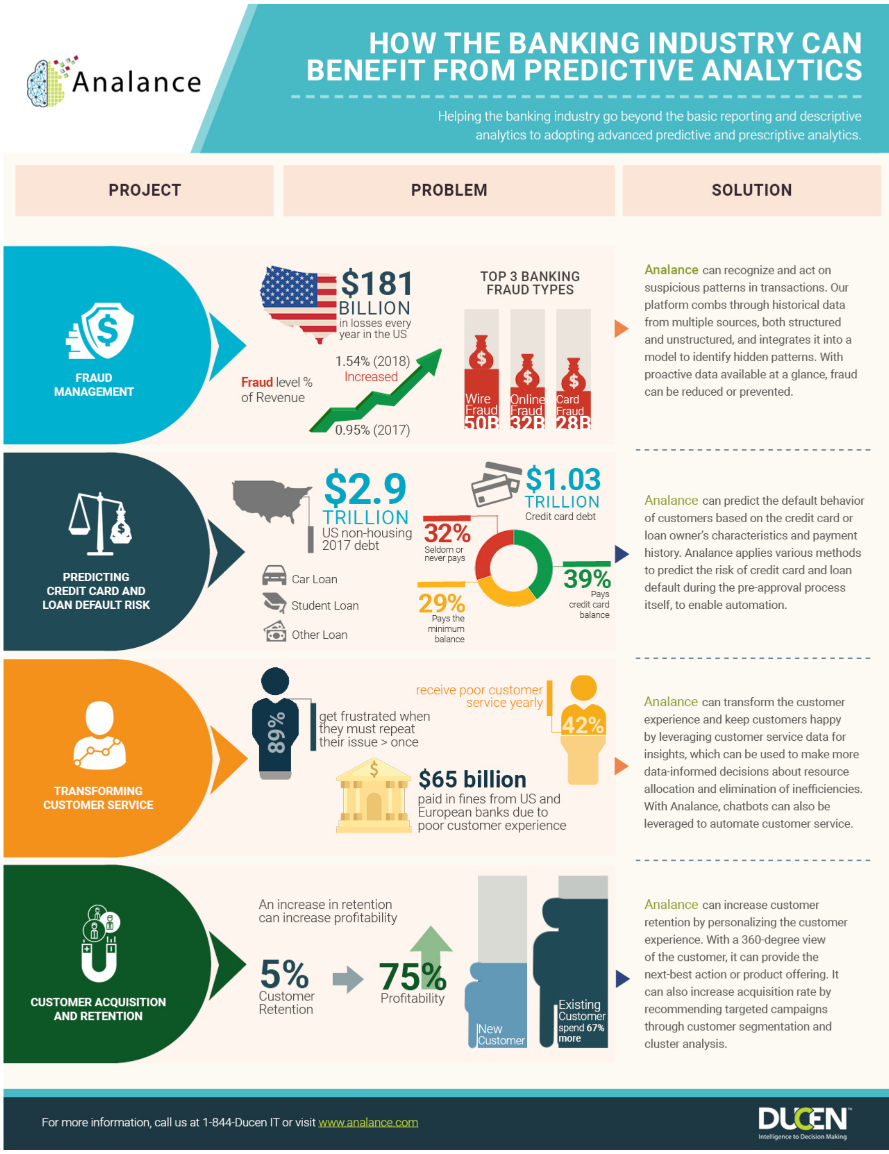
Table of Contents

About the analytics in credit card industry:

Analytics has influenced every industry based on a variety of technology platforms which collect information so, what certainly customers want is known by the service providers. The Credit Card industry is also one of these industries. There is a massive data available that can be useful in infinite ways which is enabled by credit card payment processing,

Recognition of the customer behavior:

The data from a credit card processor shows the consumer types and their business spending behaviors. Therefore, companies can develop the marketing campaigns that directly addresses consumers' behavior. In return, this helps to make better sales and the revenue undoubtedly grows.



Analance

HOW THE BANKING INDUSTRY CAN BENEFIT FROM PREDICTIVE ANALYTICS

Helping the banking industry go beyond the basic reporting and descriptive analytics to adopting advanced predictive and prescriptive analytics.

PROJECT	PROBLEM	SOLUTION
FRAUD MANAGEMENT	$181 BILLION in losses every year in the US Fraud level % of Revenue: 0.95% (2017), 1.54% (2018) Increased TOP 3 BANKING FRAUD TYPES: Wire Fraud 50B, Online Fraud 32B, Card Fraud 28B	Analance can recognize and act on suspicious patterns in transactions. Our platform combs through historical data from multiple sources, both structured and unstructured, and integrates it into a model to identify hidden patterns. With proactive data available at a glance, fraud can be reduced or prevented.
PREDICTING CREDIT CARD AND LOAN DEFAULT RISK	$2.9 TRILLION US non-housing 2017 debt Car Loan Student Loan Other Loan $1.03 TRILLION Credit card debt 32% Seldom or never pays 29% Pays the minimum balance 39% Pays credit card balance	Analance can predict the default behavior of customers based on the credit card or loan owner's characteristics and payment history. Analance applies various methods to predict the risk of credit card and loan default during the pre-approval process itself, to enable automation.
TRANSFORMING CUSTOMER SERVICE	89% get frustrated when they must repeat their issue > once 42% receive poor customer service yearly $65 billion paid in fines from US and European banks due to poor customer experience	Analance can transform the customer experience and keep customers happy by leveraging customer service data for insights, which can be used to make more data-informed decisions about resource allocation and elimination of inefficiencies. With Analance, chatbots can also be leveraged to automate customer service.
CUSTOMER ACQUISITION AND RETENTION	An increase in retention can increase profitability 5% Customer Retention → 75% Profitability New Customer Existing Customer spend 67% more	Analance can increase customer retention by personalizing the customer experience. With a 360-degree view of the customer, it can provide the next-best action or product offering. It can also increase acquisition rate by recommending targeted campaigns through customer segmentation and cluster analysis.

For more information, call us at 1-844-Ducen IT or visit www.analance.com

DUCEN Intelligence to Decision Making

Problem Statement:

In order to effectively produce quality decisions in the modern credit card industry, knowledge must be gained through effective data analysis and modeling. Through the use of dynamic data driven decision-making tools and procedures, information can be gathered to successfully evaluate all aspects of credit card operations. PSPD Bank has banking operations in more than 50 countries across the globe. Mr. Jim Watson, CEO, wants to evaluate areas of bankruptcy, fraud, and collections, respond to customer requests for help with proactive offers and service.

In this project, I will focus on the effect of the credit card limit on the credit card payments as well as customer spending. Predicting the patterns of spending and payment behaviors of customers is important for banks to provide more appealing offers to the customers, and to respond to customer requests more effectively. Although there are many factors that might be accounted for predicting customers' spending and payment behaviors, in this project, I will focus on the effect of credit card spending limit. Understanding the role of credit card spending limit will give a valuable input to PSPD bank to see how different amounts of limit affect how much customer spends and how much they pay their credit card debts. Furthermore, illustrating the relationship between credit card spending limit and the payment and spending behaviors of customers will provide evidences for banks to consider in determining the credit card limits for customers from different backgrounds and different spending and payment behaviors.

In analyzing this relationship, I will first look at bilateral correlation between credit card spending limit and credit card payment as well as between the spending limit and the amount of spending with the credit card. After doing the correlational analyses, I will develop machine-learning models. Machine-learning models will help PSD bank to correctly predict how credit card spending limit affects the customer spending and payment behaviors. The analyses will reveal important recommendations that PSPD bank and its CEO might want to consider.

About the Dataset:

The data for the analyses were acquired from Kaggle.com - Credit Card Exploratory Data Analysis. The data contain three separate files which are ;

- Customer Acquisition: At the time of card issuing, company maintains the details of customers. It has 100 observations and 8 variables including No, Customer, Age,City , Product, Li mit, Company, Segment.

- Spend (Transaction data): Credit card spending for each customer. It has 1500 observations and 5 variables including Sl No, Customer, Month, Type, Amount.

- Repayment: Credit card Payment done by customer. It has 1523 observations and 3 variables including Sl No, Customer, Month, Amount.

The dataset used for this project can be acquired using the following link: https://www.kaggle.com/darpan25bajaj/credit-card-exploratory-data-analysis#Repayment.csv

Data Preprocessing

Data preprocessing is a crucial step to be taken for a data analysis since the quality of data affects the quality of the models in predicting the concepts we attempt to explain. Hence, before jumping into performing analysis and running models, several different data preprocessing steps have been taken in order to inspect the data and look for the missing values. These steps are also important in terms of cleaning the data to make it ready for further analysis. The following steps have been taken in this section to make the data ready before feeding it into the machine learning models:

Steps taken to get the data ready for further analysis:

1. All the packages necessary for carrying out the analyses throughout this project have been imported.
2. "Credit Card Exploratory Data Analysis" dataset was read into a data-frame.
3. A general insight about the dataset was developed through checking the head, shape, info, and description of the data-frame. A preliminary analysis of the head and info of the data-frame showed that there are some columns with 23 numbers of missing values. Checking the description of the data-frame revealed that the numerical attributes looks normal (within accepted ranges) but there are some outliers were detected and they are managed. Unique values of id attribute were counted to see how many unique individuals we are dealing with.

The following section includes the results of the various exploratory data analyses These analyses have been performed to uncover the relationships between variables.

Exploratory Data Analysis (EDA)

This section focuses on exploratory data analysis which allows us to understand relationships between the variables before moving onto more complex data analysis.

I took a glance of the data in the previous section and will explore the data with the help of graphs in order to illustrate the relationships between variables. The amount of customer spending and the amount of credit card repayment are my target variables in this project. In predicting these two features, I will focus on the effect of credit card spending limit. The following graphs have been produced to explore the relationship between these three features as well as some other important features in the data.

1. Making estimations on the dataset
    - Estimation of Customer Age Mean
    - Estimation of Spend Amount Mean
    - Estimation of Repayment Amount Mean
2. Creating the following summaries
    - How many distinct customers exist?
    - How many distinct categories exist?
    - What is the average monthly spending by customers?
    - What is the average monthly repayment by customers?
    - Which city is having maximum spending?
    - Which age group is spending more money?
3. Distribution of the City Wise Spend on each Product by Year
4. Distribution of the Segment Wise Spend
5. Distribution of the Segment Wise Repayment
6. Create graphs for
    - Seasonal comparison of total spending, city wise
    - Comparison of seasonal spending for each product
    - Comparison histogram of log10 spending amount and repayment amount
7. Correlation Between Features

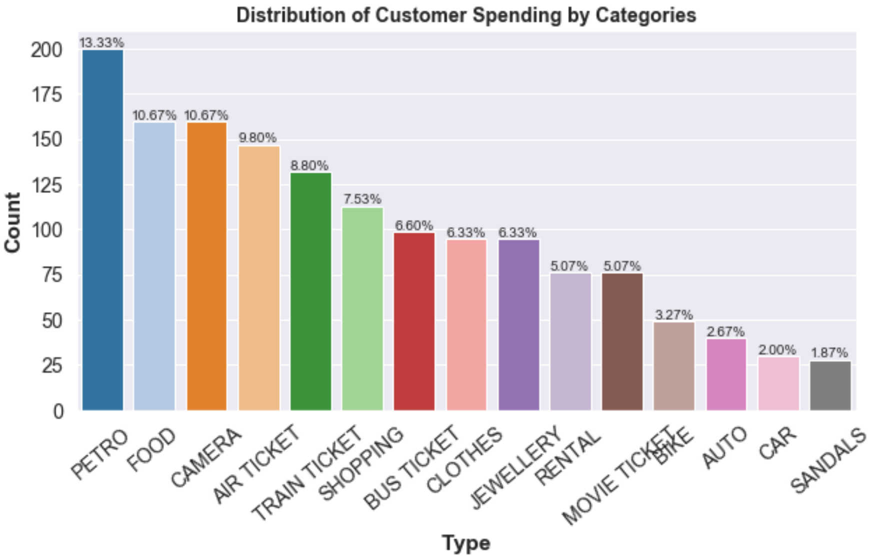
Making estimations on the dataset

- *Customers age mean: 48.4*
- *Spend amount mean : 159944.32*
- *Repayment amount mean : 163321.57*

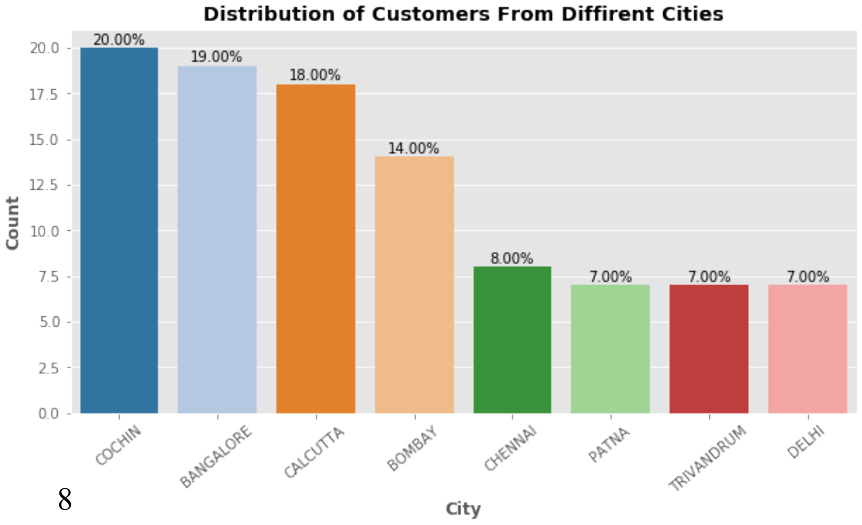
Creating the following summaries

- Distinct Customers : Number of distinct customers is 100 .
- Distinct Categories:

# Customers from Different Spend Type:
This graph illustrates the distribution of overall customer spending based on different categories. As the graph shows, the highest amount of spending has been made over petroleum products. Food and come next as 10.67 percent of all customer spending has been made per each. The lowest amount of spending has been made on sandals. Only 1.87 percent of overall spending has been made over this product.

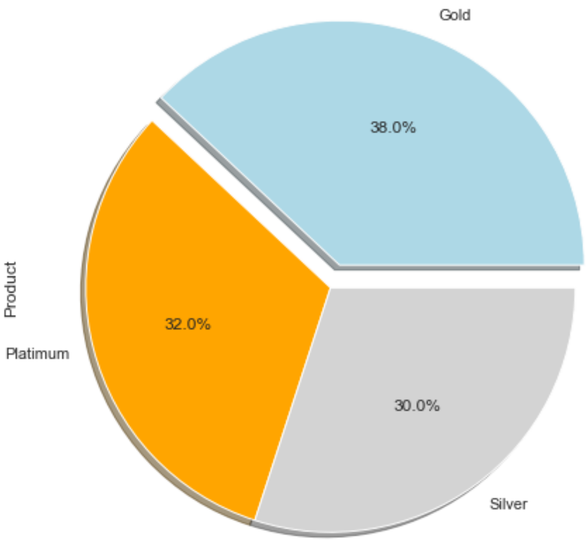


# Customers from Different Cities:
This histogram graph demonstrates that Cochin has the highest number of customers in the country while the least amount of customers using credit cards are reported in Delhi.



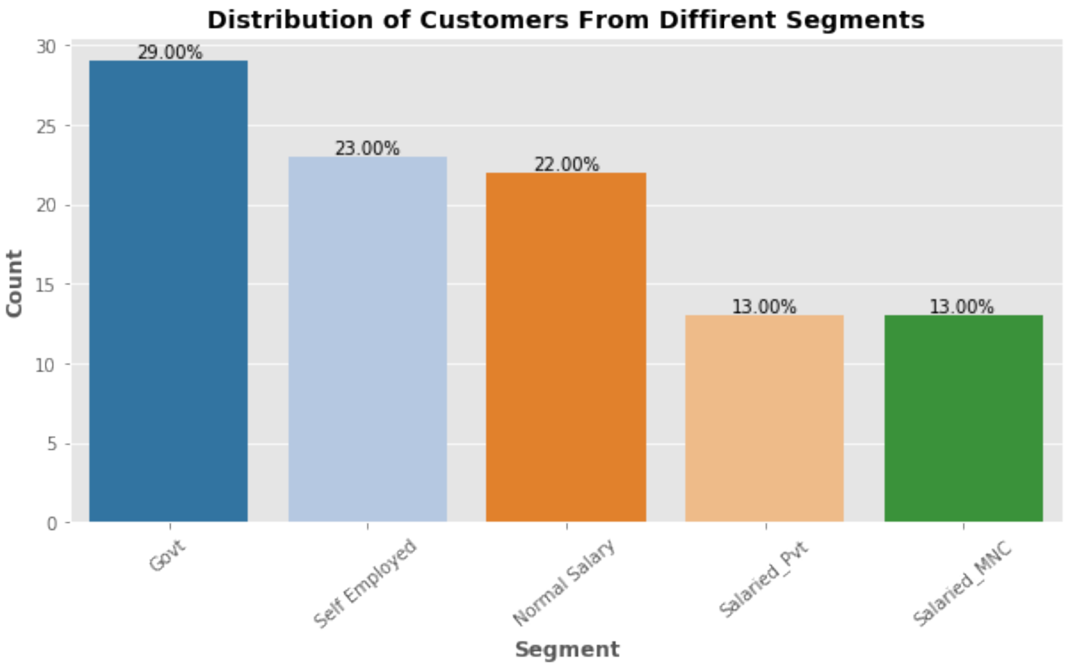
# Customers from different products:

This pie chart shows that there is not a very significant difference on the usage of different credit card products in percentage. It reports that the number of customers using Gold credit cards is higher than those using Silver and Platinum products.



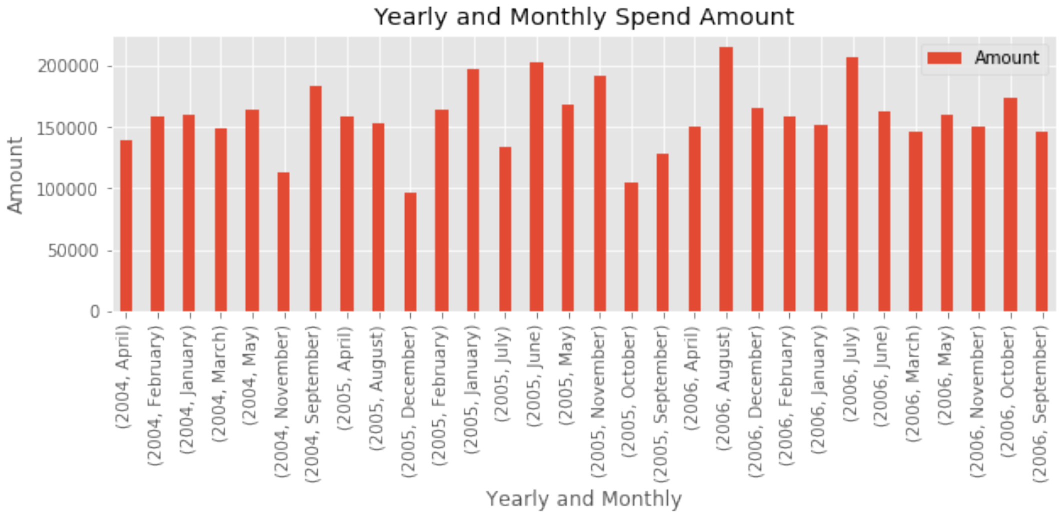
# Customers from different segments:

This bar graph illustrates economic status of the customers in the sample. It shows that the number of customers working for government (29%) is higher than people from other occupation groups.



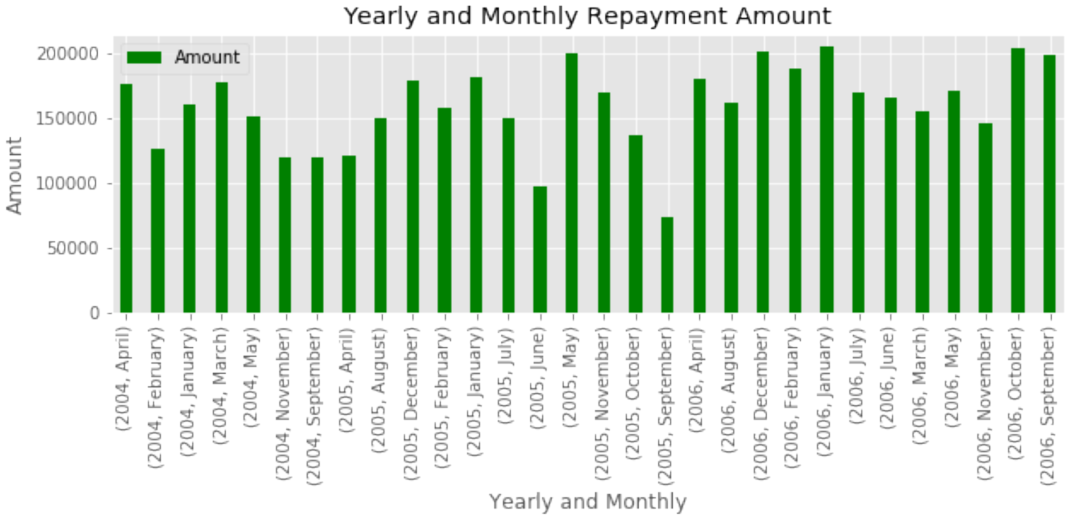
- The average monthly spending by customers:

The above red bar graph demonstrates the two-year distribution of customer spending. The time range covers from April 2004 to September 2006. It suggests that the highest amount of customer spending is reported on August 2006. The least amount of customer spending is on December 2005. The graph below shows the same two-year distribution for repayment by customers. It demonstrates that the highest amount of repayments is reported on October 2006 and January 2006. The least amount of repayment by customers is reported on September 2005. Thus, the dates of the highest and lowest customer spending do not match with the dates of the highest and lowest repayment by customers.



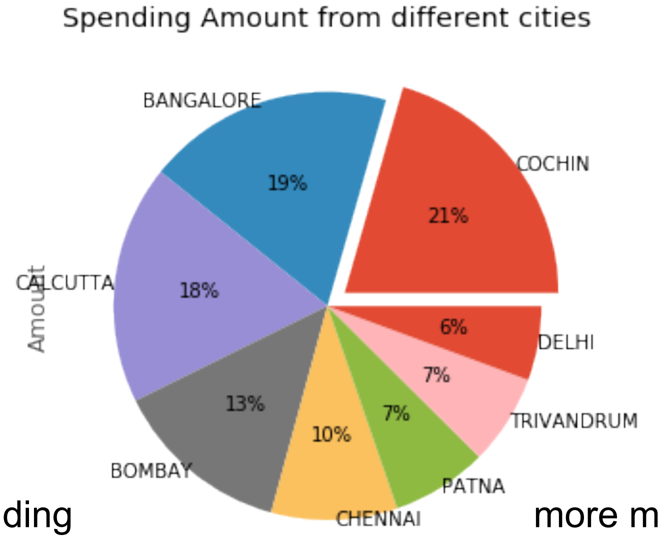
- The average monthly repayment by customers:

The below green graph demonstrates yearly and monthly distribution of credit card payments. It demonstrates that the highest amounts of credit card payments have been made in October 2006 and January 2006. The lowest amount of payment is reported on September 2005. The evidence in this graph and the previous graph suggest that we don't see an overlap in terms of the dates of highest credit card payments and highest customer spending.



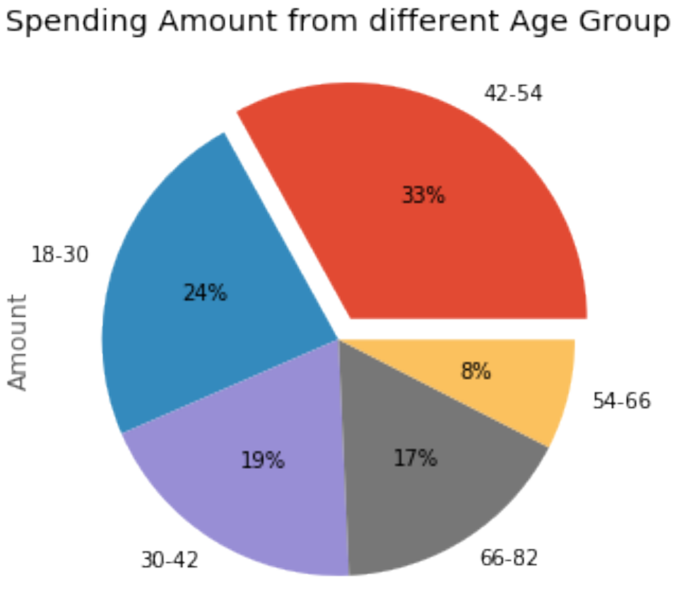
• Cities Spending Ratio :

In the pie chart below, I have illustrated the amount of customer spending in different cities. It reports the highest customer spending is produced in Cochin, which is not surprising because it includes the largest amounts of customers in the sample.

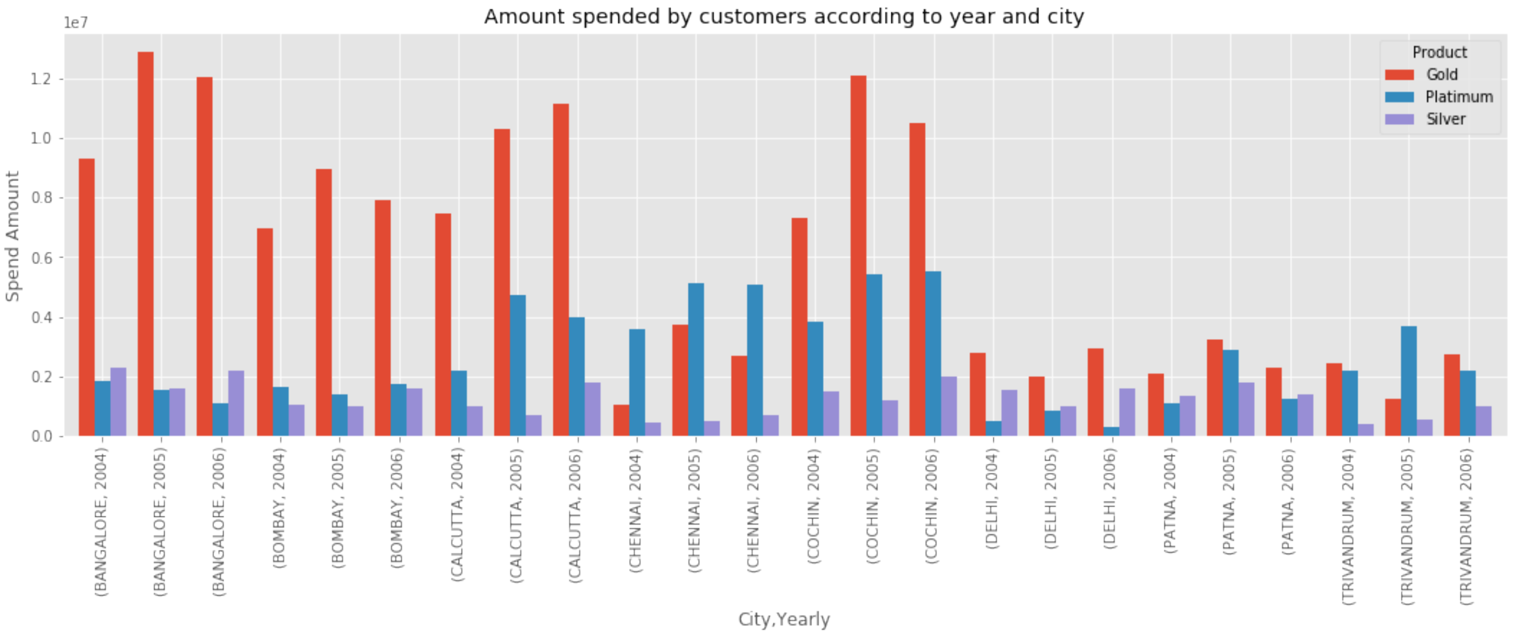


• Age group is spending more money:

From the pie chart shown below we can say that age group 42 - 54 is spending more money than people in other age groups.

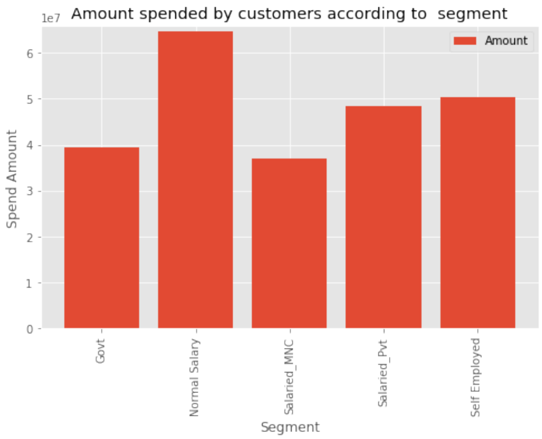


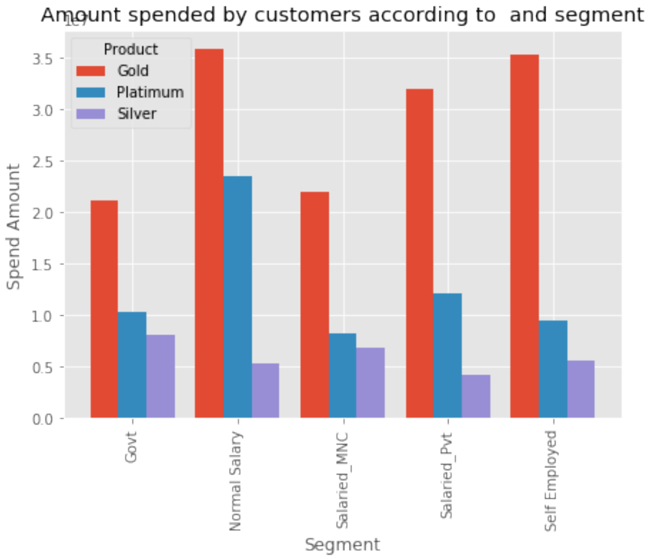
<u>Distribution of the City Wise Spending on each Product by Year</u>



This graph above demonstrates that customers with gold credit cards make higher spending than those with platinum and silver. The only two cities where customers with platinum card has made more spending than those with gold cards are Trivandrum, in 2005, and Chennai in 2004. It also reports that the highest amount of customer spending has been made in Bangalore in 2005.

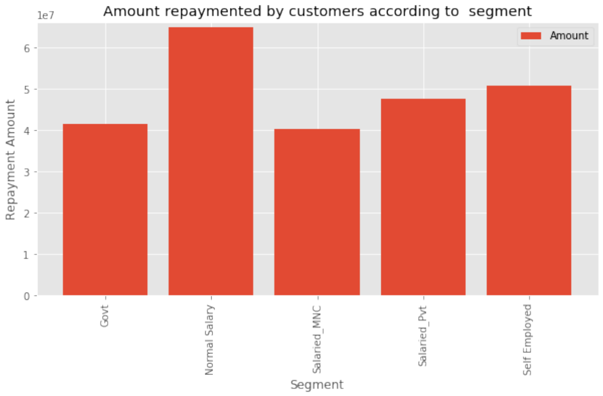
Distribution of the Segment Wise Spending:

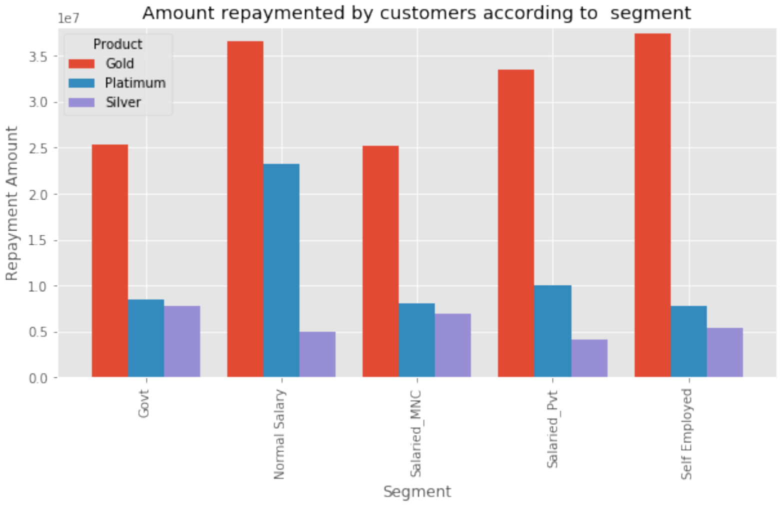




The two graphs above show the distribution of customer spending based on different economic segments. Both graphs suggest that people with normal salary are making the highest spending compared to customers in the other economic segments, particularly those having gold and platinum credit cards. However, the amount of spending made by the customers at the normal salary segment who use silver credit cards is not at the highest level, compared to customers at other economic segments who use silver cards. The graph on the left suggests that customers who work in government jobs and use silver cards make higher amounts of spending than those with silver cards at other economic segments.

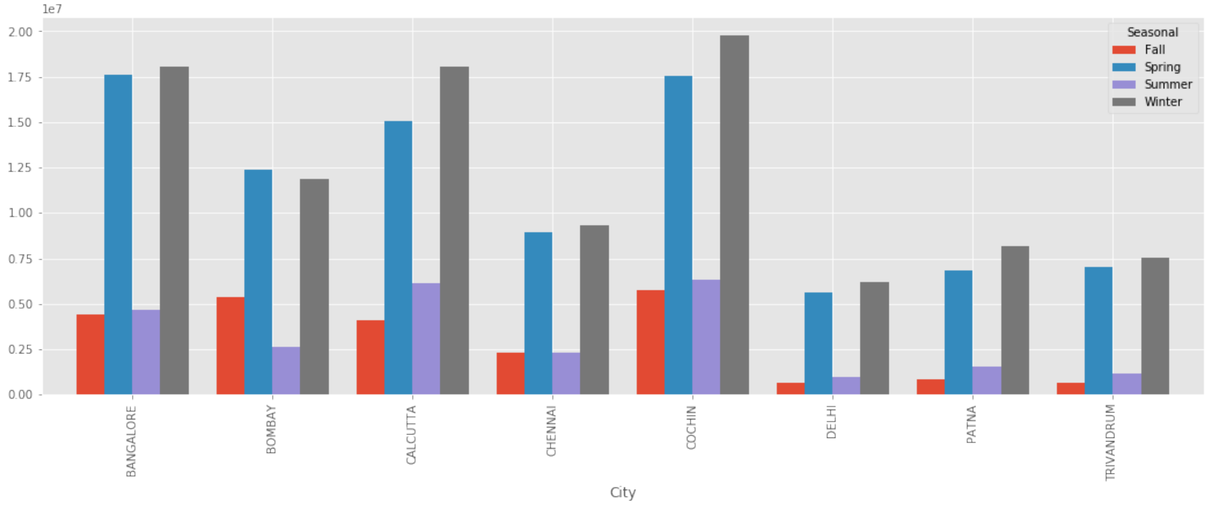
Distribution of the Segment Wise Repayment:





The two graphs above illustrate the distribution of credit card payments by economic segments. Consistent with the previous two graphs demonstrating the distribution of customer spending, customers with gold and platinum cards who are at normal salary and self-employed segments make higher credit card payments than those at the other economic segments. Those who are self-employed, however, make slightly higher credit card payments than those who are at normal salary segment.

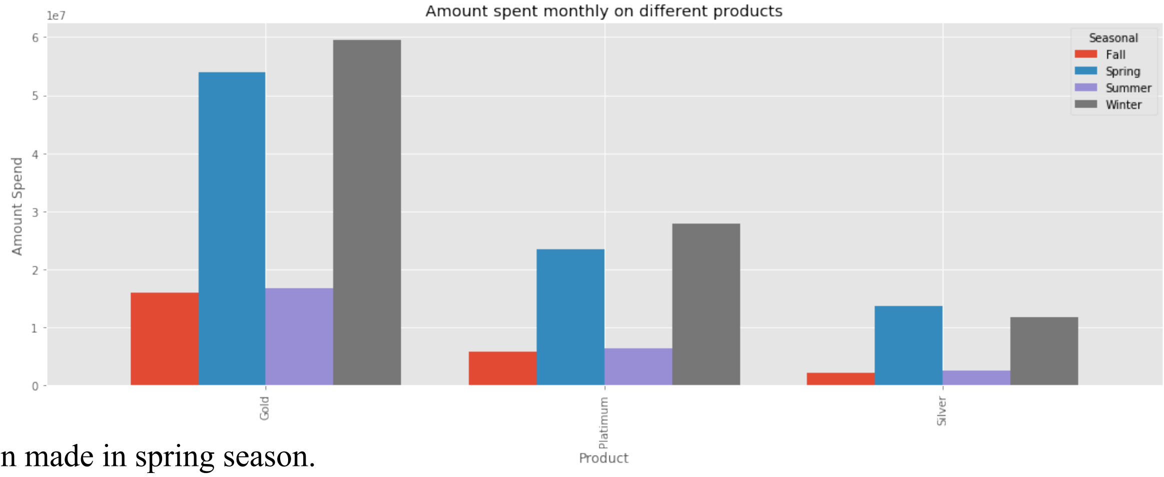
Seasonal comparison of total customer spending, city wise



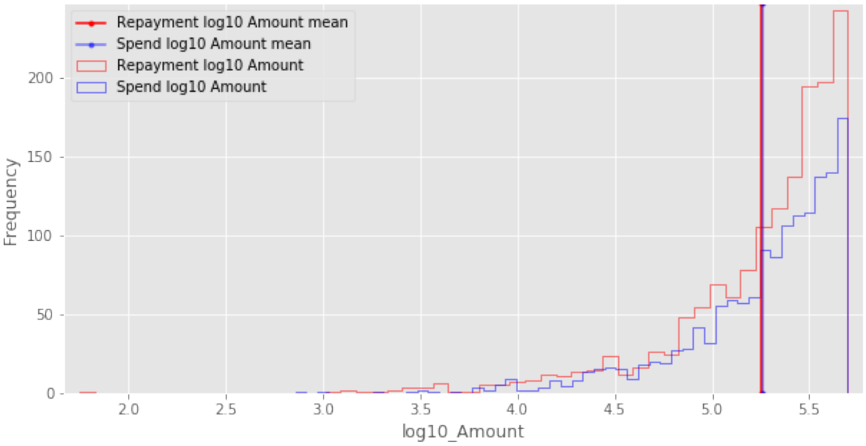
The graph above illustrates the seasonal distribution of customer spending on different cities. It reports that the winter season sees highest amount of customer spending, except for the city of Bombay where the customer spending in spring is higher than the seasons. The highest winter season spending has been made in Cochin, the highest spring season spending has been made in Cochin and Bangalore, the highest fall spending has been made in Bombay and Cochin. Finally, the highest summer spending has been made in Cochin and Calcutta.

Comparison of seasonal customer spending for each product

In the graph right, I have made a comparison across three products based on season. It shows that the highest customer spending with gold and platinum cards have been made in winter season. As for the customer spending with the silver card, the highest spending has been made in spring season.



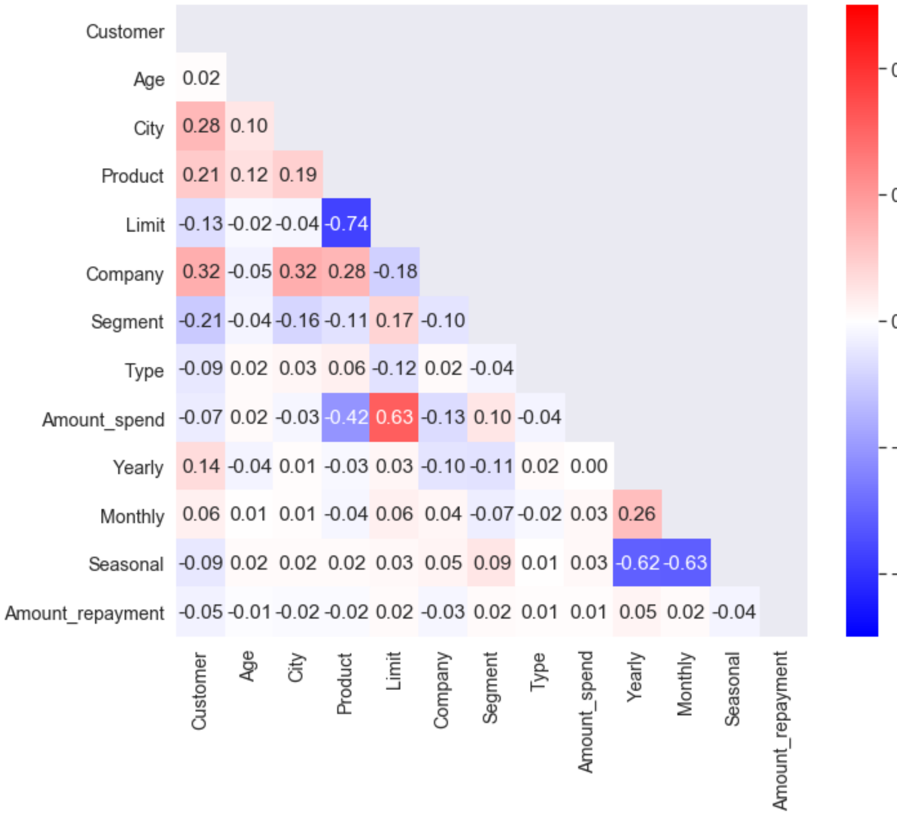
Comparison histogram of log10 spend amount and repayment amount:

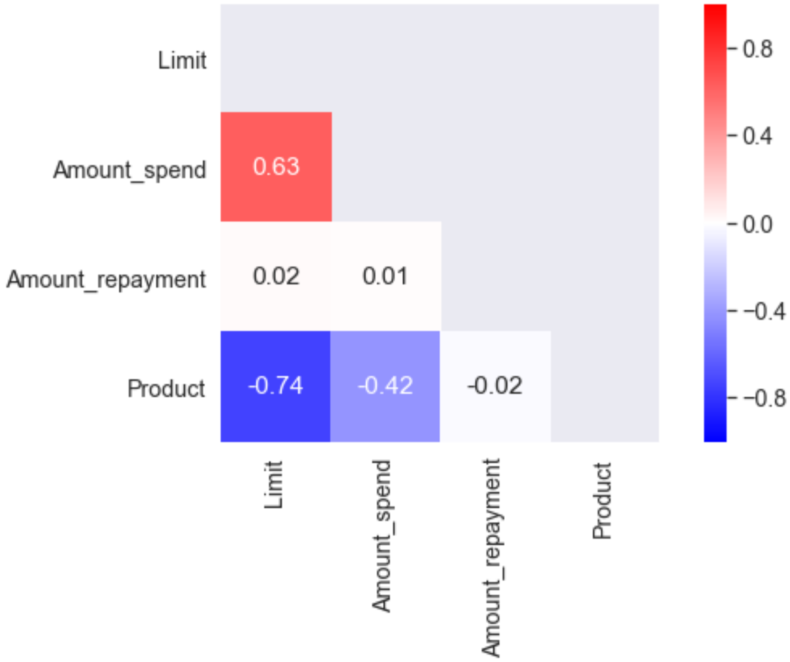


In the graph left, I have demonstrated the changes in the amounts of credit card payments and customer spending. The graph shows that as the customer spending increases, the amounts of credit card payments increases.

Correlation Between Features

Since the data includes many features, the entire dataset was divided into two parts to create a correlation matrix by using the heatmap that is illustrated below. Features with object data type were converted into categorical data type and the correlation matrix below includes those features with categorical codes.The first heatmap illustrates the relationships between the variables. While the red color indicates a positive correlation between the two variables, blue color indicates negative correlation. The magnitudes of correlations have also been shown on the heatmap. In the second heatmap, I looked at only the variables of interest in my project, which are credit card spending limit, customer spending, and credit card payments.





The second heatmap demonstrates that there is a positive and significant correlation between credit card limit and customer spending. The correlation coefficient is reported 0.63. However, the heatmap below shows that credit card payment isn't significantly correlated with credit card limit (p=0.02) and

Feature Selection

Feature Selection for Object Data Type

The descriptive statistics of the features on the object data type demonstrate that several features should be dropped because they have a large number of unique values that would lead to memory error. The features that are highly correlated with each other have also been dropped since including just one of those features would be sufficient to feed my machine learning models. In addition to these features dropped from the data, some features were also dropped from the dataset because the information from these features do not have a significant value for further analyses. The features I dropped are company name, customer name, and the month of customer spending and the month of customer repayment.

One Hot Encoding

Now I have 7 features on the object data type after I drop the features having low variance or highly correlated with one or multiple features in the data. I have converted these 9 features into numerical data type since categorical variables won't work in the algorithms of machine learning. In order to convert these features, it would be possible to use "label encoder" or "one hot encoder". "Label encoder" wouldn't be the right one to use because most features of mine on the object data type include multiple categories. Therefore, I have used "one hot encoder" in converting these features into numerical type.

FEATURE NAME	DESCRIPTION
AGE	The age of the customer
LIMIT	The credit card spending limit
AMOUNT SPENT	the amount of the customer's spending with the credit card
AMOUNT REPAYMENT	The amount of credit card payments made by the customer
CITY	The city in which the customer lives
SEGMENT	Occupation type
TYPE	The name of the purchase
SEASONAL	The season of the spending
AGE GROUP	What age group the customer is located
YEARLY	The year when the customer spending or credit card payment happened
MONTHLY	The month when the customer spending or credit card payment happened

Now, we selected our features and got the data ready to build our machine learning models.

Model Building / Machine Learning

The central goal of this project is to correctly identify and predict customer spending by utilizing various machine learning models.

Model Fitting:

For model fitting, I have used the following machine learning algorithms in order to predict customer spending.

1) Linear regression

2) Random forest regressor

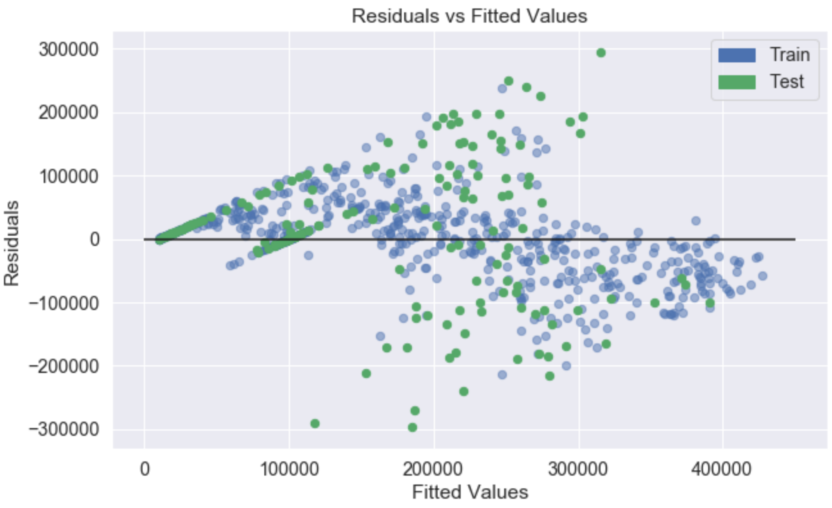
3) Optimized random forest regressor with GridSearchCV

I have split the data into train (80%) and the test (20%) sets. Then, I have scaled the train and test sets by using Scikit-learn's StandardScaler that would help me to standardize features by excluding the mean and scaling the sets to unit variance. Only continuous numeric features have been subjected to such a scaling. Categorical variables were excluded in doing the scaling. I have first done linear regression and the model reveals that R-square for train data is 0.44 whereas the R-square for test data is 0.36. When I use random forest regressor, R-square for train data is 0.89 while R-square for test data is 0.35, suggesting that it explains 89 percent of variance in the train data and 35 percent of the test data. These numbers tell us that the model might be overfitted with the training data. To address this issue, I use GridSearchCV for random forest regressor to optimize these numbers. GridSearch is useful in the sense that it tries all different combination of hyperparameters in order to find the best score. Cross validation is important because it is useful in eliminating bias in the model's parameters.

The performance statistics may appear extreme because of the random nature of the data that is helpful for splitting the data into training and test sets. To mitigate this issue, the cross validation is used since cross-validation allows for using multiple samples from the dataset.

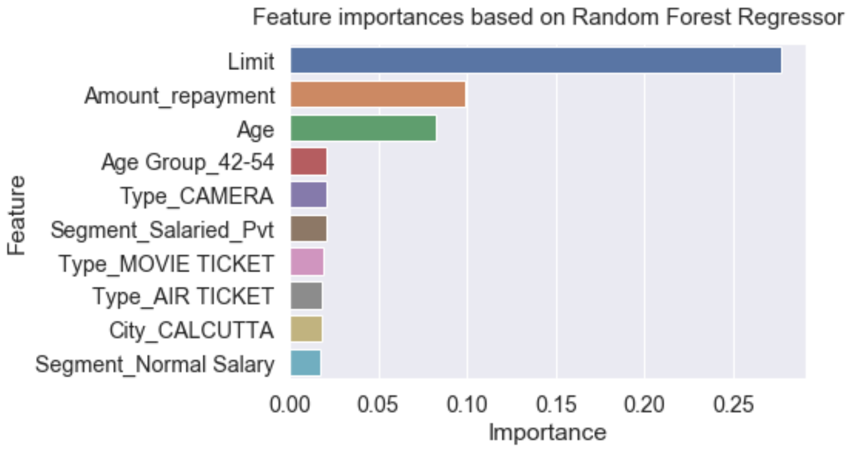
I have conducted grid-search and cross-validation to enhance prediction performance of my model and to enhance generalizability of our model. The results reveal that R-square has turned out to be 0.86 for the training data and 0.45 for test data.

To conclude, random forest has provided a better score than linear regression in terms of explaining the variance. I have further improved the score by doing GridSearch and cross-validation.



Feature Importance

The random forest regressor could tell me how useful a feature is in predicting an outcome because it is built upon a decision tree. I have presented below the top 10 features with highest importance in predicting customer spending. The graph below shows that the following features are most important ones: credit card spending limit, the amount of credit card payment, age of the customer, the age group ranging from 42 to 54, the customer purchase of camera, air ticket, movie ticket, customers living in Calcutta, customers with salaried private and those with normal salary. Credit card spending limit has the largest importance, and the magnitude of its importance is significantly higher than other 9 features.



Limitation

As the R-square scores suggest, there is a significant difference between R-squares for train data and test data. This is a major limitation of this project. There might be some reasons explaining this limitation. For one thing, the sample includes only 100 people. Including higher number of people would change the results. Secondly, I don’t have info about the population of each city. Including information about the population would give me a better understanding of why customer spending and credit card payment amounts are significantly higher than other cities, thereby helping me further in predicting customer spending. Finally, beyond the individual-level factors, systemic factors would be taken into account to better understand people’s spending patterns. For instance, a financial crisis or some sort of an economic shock would definitely affect the levels of customer spending. So, if I have accessed data on these information, the prediction performance would increase and the difference between trained data and test data in terms of R-square would decrease.

Business Recommendations

I finalize the project with the following recommendations:

1. Exploratory findings from the data demonstrate some patterns that may help PSPD bank to increase its profit and reach more customers. For instance, customers spend more on petroleum products. In addition, people with normal salary or self-employed spend more than other occupation groups. It is also clear that the amount of customer spending with gold credit cards is higher than those made with other credit card types. Finally, people whose age on the range between 42 to 54. Based on these findings, I would recommend that PSPD would develop deals on petroleum products for more customer spending. Also, PSPD would encourage self-employed people or people with normal salary as well as people whose age are in the range between 42 to 54 because these people spend more. Finally, since credit card spending amount with gold card is higher than other credit cards, PSPD may want to develop some deals to reach and encourage people to apply for gold credit cards.

2. The machine learning models that I have developed here successfully predict customer spending with credit cards. These algorithms can help the PSPD to predict customer spending patterns.

3. The project shows that credit card limit is the most important feature in predicting customer spending. Therefore, PSPD may want to consider increasing credit card limit for certain customers to encourage more spending by customers. Specifically, the exploratory findings on credit card repayment demonstrate that self-employed people and people with normal salary make more credit card payments than people from different occupation groups. Therefore, PSPD may want to increase credit card limit of self-employed people and people with normal salary.

Conclusion

The central goal of this project was to predict customer spending with credit card. The processes of data wrangling, exploratory data analysis, feature selection, and model building were performed on the PSPD Bank's dataset. The insights gained as part of these processes were highlighted and several business recommendations were provided for the PSPD Bank.

The code associated with this report is available at: