Credit Card Spend Prediction Report

Rabia Tariq

August 2021

Data Sourced from: Kaggle.com

Introduction:

Credit cards are great, convenient method of making payments, easier to carry than cash, helps you build a credit rating and are there if you need emergency cash. Banks provide these credit cards to an individual and determine the credit limit for each card based on certain factors. People are willing to spend more when paying with a credit card instead of cash. We've gathered data by conducting a survey of 5000 customers of a bank. This includes data about age, gender, income, job, card spend to name a few. It contains information about customer's primary and secondary credit cards. Using this, we can get an idea of the customers' spending habits.

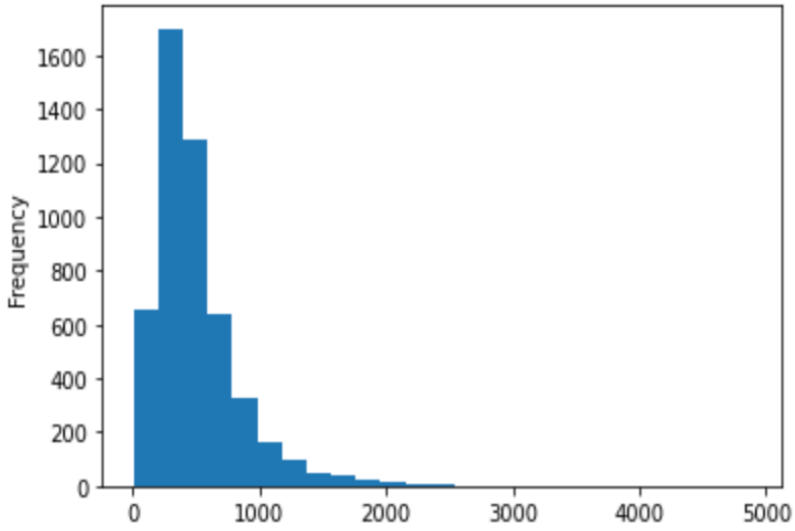
Problem:

Banks like to understand what the factors driving credit card spend are. The bank wants to use these insights to calculate credit limit. Lenders like to understand the factors driving the credit card spend. How can we use the data to understand what is driving the total spend (Primary and Secondary cards), and given those factors predicting the limit of their credit card spent.

Data Wrangling:

The data originally came as two csv files. One with all the customer information and the other was a detailed dictionary explaining what each of the feature means. The data was encoded in the numerical format to reduce the size of the data, however some of the variables are categorical. This dataset had a lot of variables, most of them being categorical but encoded as numerical. The data was relatively clean, only a few variables having more than 25% of the data missing. There were some variables that had very few values missing, so I imputed them with their mean values. With those variables, as a lot of values were missing so I handled that by dropping those variables. There were also some outliers present in the data, so I created summary statistics and clipped the values that were outside my specified interval to the interval edges.

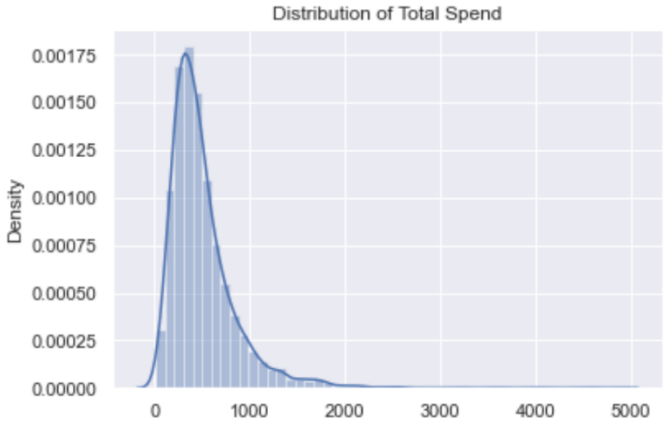
As my target variable was the spend of the credit card, I combined the primary and secondary credit card spend and created my target variable. Then I got my first look at the distribution of my target variable.



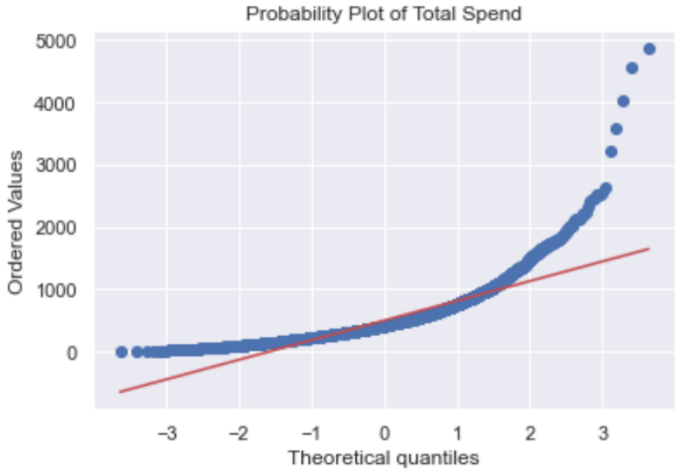
After I was able to get somewhat of a cleaned data, I created dummy variables for the categorical features.

Exploratory Data Analysis

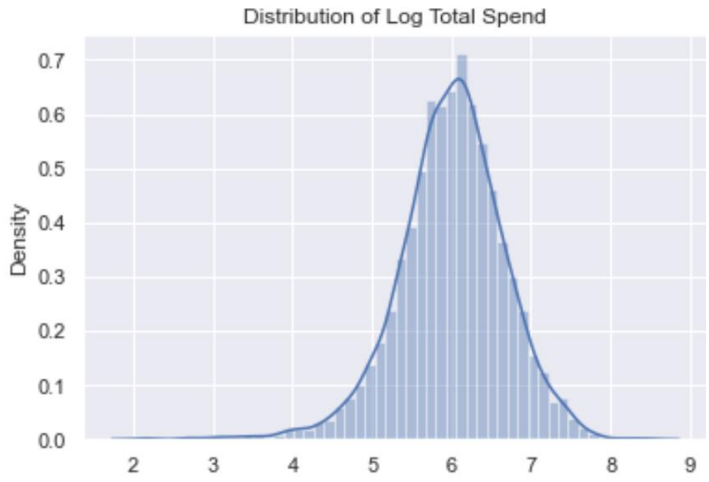
I started off by first looking at the distribution of my target variable again. Regression assumes that the distribution of our dependent variable is normal.



Here we can see our target variable is right skewed. Next, I wanted to look at distribution of our dependent variable against the theoretical normal distribution with the help of probability plot, so that we can graphically see and assess that whether or not our target variable follows the normal distribution.



As we can see from the plot our data was not distributed normally. So, the next step I did was use the log of total spend and again look at the distribution to see how it changed. I previously created the column total_spend_ln which was created using the log of the target variable. So, the distribution of the log of target variable looked like the following.



As can be seen from the graph, by taking the log the skewness of the dependent variable is greatly reduced. Again, I graphically looked at the distribution of the log variable against the theoretical normal distribution using the probability plot.

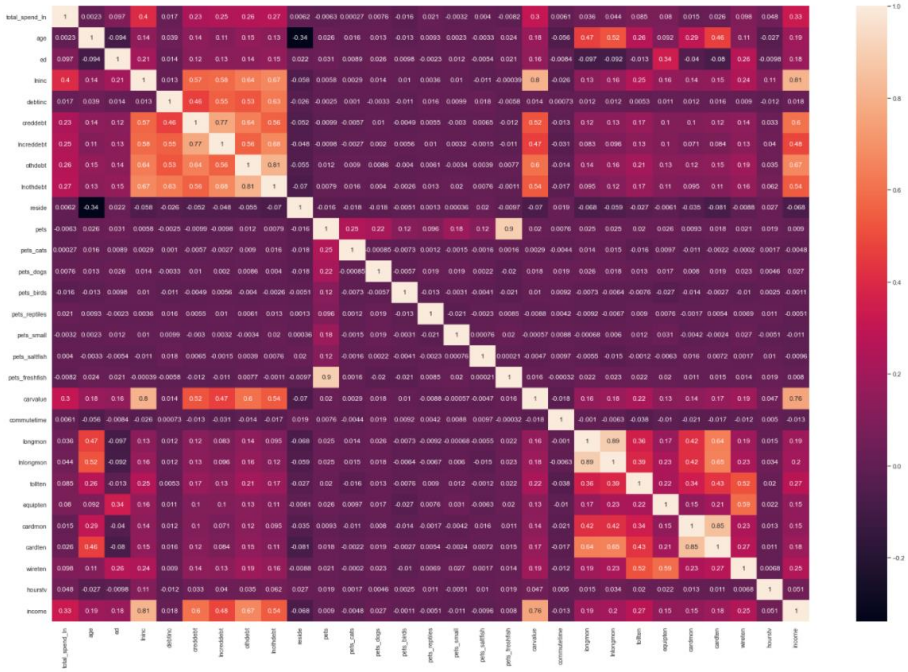
As can be seen from the above graph, the log variable distribution is now a lot closer to the theoretical normal distribution. I learned from the analysis of the distribution of the target variables that total_spend is right skewed. So, taking the log of it reduces the skewedness. So, I dropped the total_spend column from the dataset and chose total_spend_ln to be my target variable.

Next, I wanted to see how each variable correlated with my target variable and also with each other. For that I created a heatmap to represent the data in 2-dimensional form and provide a

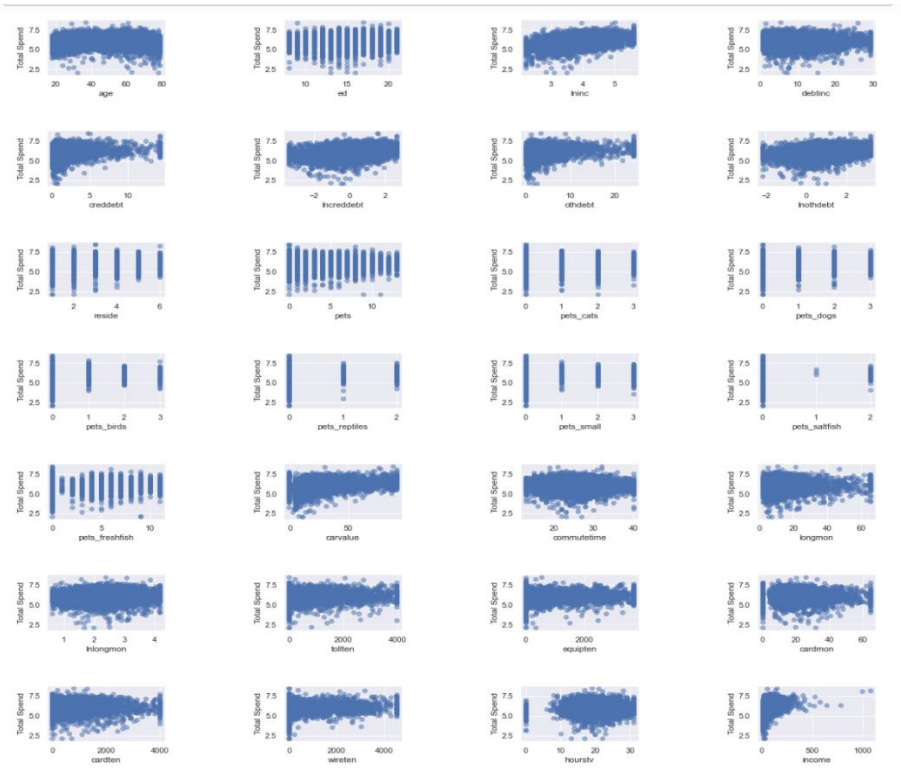
colored visual summary of information. It will help us find which variables have the highest correlation with each other and our target variable.

Here we can see some correlations between the different variables. We are more interested in the correlations with our dependent variable total_spend_ln. Some of the correlations and important features we can see which will be useful in our model to make predictions of our target variable are:

- income
- carvalue
- lnothdebt
- lncreddebt
- lninc



Correlations, particularly viewing them together as a heatmap, can be a great first pass at identifying patterns. But correlation can mask relationships between two variables. So, I created a series of scatterplots to really dive into how total spend varies with other numeric features.



In the scatterplots we see some of the correlations were clearly picking up on. There's a positive correlation with lninc. Our target variable seems to have the strongest correlation with **lninc**. **lncreddebt** and **lnothdebt** appear quite similar but also useful. **carvalue** also has correlation with our target variable and seems to be useful.

For our target feature total_spend_ln, there are a few correlations observed through heatmap and scatter plots. The correlation with car value means that people who have cars with greater vehicle sticker prices, tend to spend more on their credit cards monthly. Similarly people with higher income spend more monthly on their credit cards.

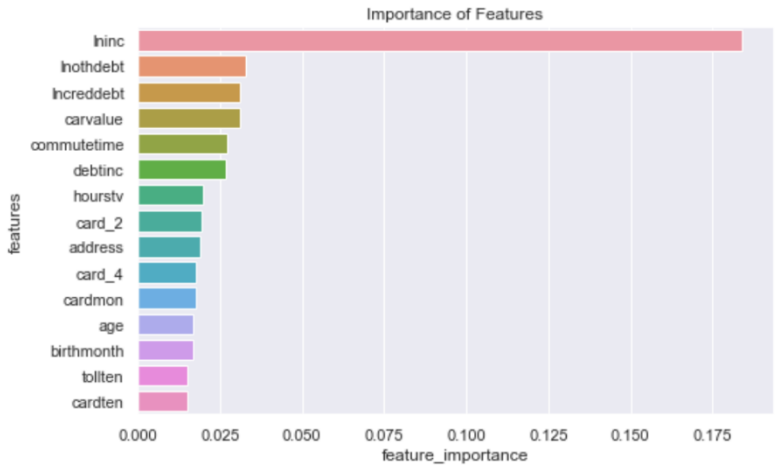
Preprocessing and Training

Until this point, we've treated our data as a single entity. In machine learning, when we train our model on all of our data, we end up with no data set aside to evaluate model performance. We could keep making more and more complex models that fit the data better and better and not realize we were overfitting to that one set of samples. By partitioning the data into training and testing splits, without letting a model (or missing-value imputation) learn anything about the test split, we have a somewhat independent assessment of how your model might perform in the future.

Random forest has several hyperparameters that can be explored, however here I limited it to exploring some different values for the number of trees using GridSeachCV. I used Random Forest Regressor to look at feature importance and see which are most important for our model to make predictions.

The top 4 features were:

- lninc
- lnothdebt
- lncreddebt
- carvalue



At this point, we still had a lot of features in the dataset. So, the next step I did was use Recursive Feature Elimination for selecting the features. After getting the list of selected features, I had to deal with multicollinearity. In regression, "multicollinearity" refers to predictors that are correlated with other predictors. Multicollinearity occurs when your model includes multiple factors that are correlated not just to your response variable, but also to each other. In other words, it results when you have factors that are a bit redundant. Multicollinearity increases the standard errors of the coefficients. By overinflating the standard errors, multicollinearity makes some variables statistically insignificant when they should be significant. So, I dealt with it by using the Variance Inflation Factor (VIF). I checked the VIF of each feature and then dropped those with VIF above 10, because generally VIF above 10 indicates high correlation and is a cause of concern.

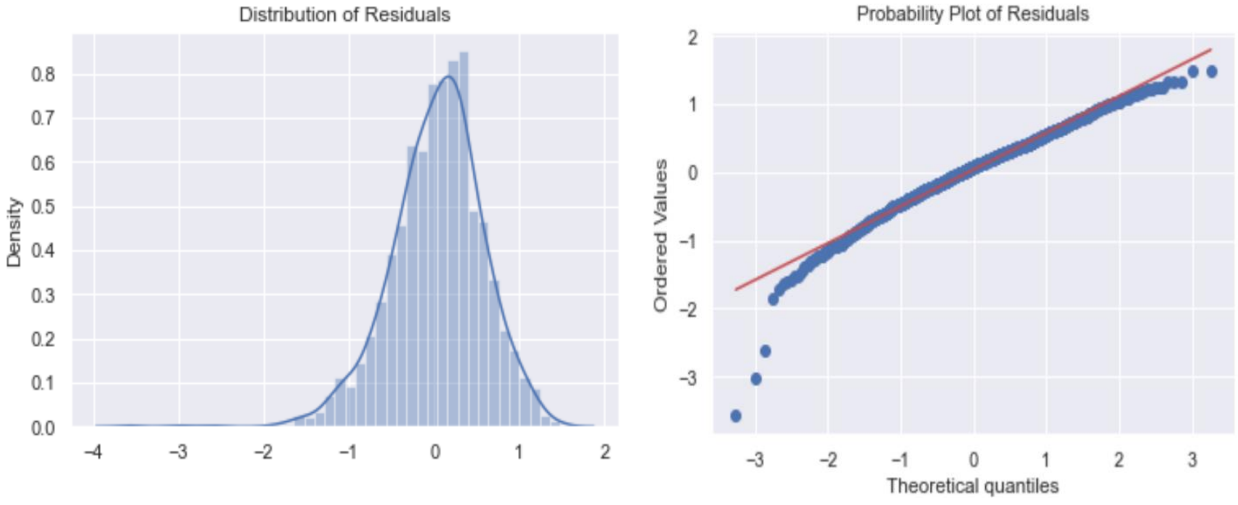
Modeling

The goal with this modeling was to predict the monthly total spend of the customers. I split my data into 75% training and 25% testing.

The first model that I created was using Ordinary Least Squares (OLS). After fitting my model on training data, I looked at their p-values. I then created a list of features with p-values greater than 0.05 and then dropped those variables from my data and then fit the model to the training data again. By doing that my MSE, MAE and RMSE decreased.

	total_spend_ln	pred_total_spend_ln
3010	5.023288	5.903262
1376	5.741399	5.823382
4368	5.634682	6.076664
713	5.528992	5.783019
206	7.793595	6.777161
...	...	...
1834	5.692755	5.865804
2833	5.690292	5.851115
3218	6.169778	5.454478
3228	7.006306	6.871852
419	5.801937	6.784364

I then looked at the distribution of the residuals. Residuals, in the context of regression models, are the difference between the observed value of the target variable and the predicted value, i.e. the error of the prediction. Now we check whether the residuals are normal. If the residuals are normally distributed, then their quantiles when plotted against quantiles of normal distribution should form a straight line.



My next model was created using Random Forest. I used GridSearchCV to tune my hyperparameters and got the best number of trees to use in my model. I then fit the model to my training data and again evaluated my model using mean squared error, mean absolute error and root mean squared error. My third model was created using XGBoost and evaluated using the same metrics. Then I compared my models using these metrics i.e., MSE, MAE and RMSE.

	MSE	MAE	RMSE
OLS	93467.195826	197.760448	305.724052
RandomForest	103425.755175	208.295764	321.598749
XGBoost	108212.096835	216.376562	328.956071

From the mean squared error, mean absolute error, and root mean squared error of our different models, it can be seen that the OLS model performs the best. So, I used OLS model to make our predictions.

Predicting Credit Card Spend

Our target variable was the log of the total spend. So, to get the predictions of the total spend, I took the inverse of our predicted values and created a table to look at the predicted and actual values of monthly total spend. Shown below is the that table with five of the actual vs predicted values.

	Actual_total_spend	Pred_total_spend
3010	151.91	366.230224
1376	311.50	338.113679
4368	279.97	435.573564
713	251.89	324.737996
206	2425.02	877.573900

Conclusion

We have done the data wrangling, exploratory data analysis, preprocessing and training, and finally built some models. Our objective was to find the important features which determine what drives the bank customers' monthly spending on their credit cards. Also, we wanted to create the model which predicts that spend. We have built our model and predicted the total spend of the customers. All this work is important because it can help banks determine the credit limit of customers. As banks look at different factors in determining the credit limit of an individual, one of them being their spending history and spending habits. So, the banks can use the predictive model to help inn determining the credit limit of the customers in future.