Credit card spend Prediction and Identifying the key drivers

Objective Finding the Credit limit for the card holders

Business Problem

- Find the key factors driving the Spend (Primary card + Secondary card)
- Predict credit limit for new customers based on the credit card spend

Data Availability

- Customer demographics, customer spend details, card type details (reason for using, type of card, card benefits), customer personal details such as the (job type, commute, properties own etc.)
- About 130 Features for 5000 customers

Approach

1 Pre-modelling and Data audit

- 1.1 Pandas profiling
  Initial identification of categorical and numerical variables, variables with missing values, high correlation, high zeros, duplicate rows
- 1.2 Identifying the categorical variables
  As per the given data dictionary the categorical variables are identified.
- 1.3 Finding the variables which can be considered for modelling
  By applying the Decision Tree regressor on the entire independent variables the top 20 variables which are found as importance through the feature importance attribute were selected.
  By the business understanding the few variables were also added.
- 1.4 Analyzing the spend (Dependent variable)
  The spend is represented by the two variables Primary and Secondary spend for each primary and secondary credit card respectively.
  The distribution of the spend was initially skewed and has some kurtosis but the Logarithmic transformation helps it become close to Normal distribution

2 Data Preparation

- 2.1 Outlier and missing value treatment
  The distribution of each variable was analysed with the box plot. It is found that few variables such cardspent, card2spent were having some outliers which was then treated using the 1% - 99% capping.
  The missing values for the numerical variables were imputed with the median values.
- 2.2 Determining the assumptions
  As the statistical model Linear Regression needs the dependent variable to be normally distributed, the log(spent) will be used while building the model.

3 Feature Reduction

- 3.1 Featuring reduction considering only Xs
  - Correlation between Xs
    The correlation matrix was created considering only the independent variables and the highly correlated variables were noted and chosen for elimination.

- Principal Component Analysis (PCA)

  The PCA was applied on the independent variables it is identified that about 20 Principal Components were able to explain 80% of the variance of the overall set of independent variables.

  The Factor loading matrix was then created and the variables which have high correlation with the initial PCs were selected.

- Variance Inflation Factor (VIF)

  The VIF was performed on the independent variables set and the variables having the VIF value above 5 were excluded and the lowest ones were selected.

- 3.2 Feature reduction based on X and Y

  The independent variables are the derived variable of the sum of Primary card spent and Secondary card spent.
  For each method 15 variables were selected.

  - Recursive feature elimination (RFE)
  - F Regression
  - Select K-best
  - Decision Tree feature importance

- 3.3 Combining the all the selected features and splitting the data into train and test

  X = The selected variables from the set of both Feature reduction methods were taken for the final model building.
  Y = log (Primary + Secondary spend)
  The train and test split was made on the above

4 Model Implementation + tuning hyper parameters + model evaluation

The hyper parameters of the machine learning models were tuned with the help of GridsearchCV by providing the set of hyper parameters grid. The best set of parameters were finalized for each of the ML model.
The cross-validation number is set as cv = 5.
The Model Evaluation metrics used are MAPE and RMSE.

- Regularized regression techniques
  - Ridge
  - Lasso
  - Elastic net

  The penalty term alpha was set to 0.01 for all the above methods.
  Out of the above three models the Ridge was giving a better result of the evaluation metrics.

- Ensemble methods
  - Random Forest

    The hyper parameters tuned were number of estimators and maximum depth. The corresponding best values are { max_depth=8, n_estimators = 350 }
  - XGBoost

    The hyper parameters tuned were number of estimators and maximum depth. The corresponding best values are { max_depth=4, n_estimators = 38 }

- Support Vector Regressor (SVR)

  The hyper parameters tuned were cost, gamma, kernel type, epsilon value. Since the linear SVM gives the best results, the non linear parameters were left unchanged.

	Ridge	**Lasso**	**ElasticNet**	**Random Forest**	**SVR**	**XGboost**
MAPE_train	0.489022	0.519883	0.503645	0.40012	0.459744	0.434494
MAPE_test	0.452357	0.477457	0.462347	0.482776	0.434113	0.430395
RMSE_train	272.84869	290.852137	283.021019	240.967713	299.924879	285.885744
RMSE_test	272.409226	290.665433	282.326541	285.516083	300.886334	292.841504

The MAPE and RMSE values were high for all models due to the fact that given data was not having a good correlation with the spend. Or the data set it missing some of the features. Although the while comparing the models it is found that the SVR and XGBoost models were performing good relatively.

5 Finding the Optimum Credit Limit for the card holders

As the objective of this business problem is to predict the spend thereby setting the Credit limit for each customer.
We assume that the Credit limit can be set as 3 times the total spends of the customer.
While the minimum Credit limit is set as 165,000 or 3 times the spend which ever is high.

6 Identifying the Key Drivers of the Spend

In order to identify the key drivers of the spend, as per the data dictionary the card type, card benefit, job categories columns were renamed for getting a better interpretation.

- Implementing the stats model Linear Regression
    - Identifying the important variables from the Model Summary
      The Linear Regression was implemented and for a few iterations the unimportant variables were removed by reading the summary report and re-implemented.
      The final model derived after neglecting the unnecessary variables
    - Math Equation (for Excel tool)
      The math equation for predicting the spend is derived from model summary as summation of intercept and the product of coefficient of each selected variable and the variable.

      This math equation can be used in the Excel is determine the spend of the customer.
- Summarizing the Key drivers of the spend (Positive and Negative drivers)

  The key drivers were identified from the model summary (here Linear regression and Ridge models were considered), which are variables having higher coefficient values.
  If the coefficient is Positive/Negative high then that variable has high Positive/Negative relationship with the spend.

7 Conclusion on Key Drivers of Spend

Positive Drivers

- Reason_2 - People who use the card for the reason "convenience" tend to spend more and hence will have high Credit limit
- Marital - If a person is married, will spend more
- Income - High income more Spend

Negative drivers

- People using the Discover and Mastercard have low Credit limit
- Female spend less
- If a person is Retired spent less
- People of Sales_and_Office_job spent less

Challenges

- Feature Selection - As the data given has lots of unnecessary variables which do not contribute to the prediction of spend it was difficult to reduce / select the important features.
- Since there was only mild correlation between the dependent and the independent variable, Obtaining the accuracy / performance of the model was a difficult task.