**Capstone Project 1**

**Predicting Credit Card consumption**

*Data from one of the leading banks is used to estimate the credit card consumption of customers.*

*Demographic data, Behaviour data and Credit consumption data was merged to make this a* ***supervised******regression*** *problem. A number of Machine learning techniques are analyzed and final solution is given to predict the credit card consumption of customers.*

* Business Objective

Understanding consumption pattern at customer level to understand customer behaviour. Customer understanding allows banks to develop marketing campaigns addressing customer needs, which in turn increases sales.

* Input

Three different types of data are provided to us

1. Customer Demographics data

* This contains features like age, gender, income, region, employment tenure etc. specific to each customer.

1. Customer Behavioral data

* Behavioral data include information on spending pattern, assets and liabilities of each customer.
* Spending for credit card and debit card is provided for 3 months (april, may, June)

1. Credit Consumption data

* This contains the **Target** variable, Average credit card consumption for next 3 months for each customer.
* Data Consolidation
* All three datasets i.e., Customer Demographics Customer Behavioral and Credit Consumption were merged on the basis of unique **Customer ID** feature.
* Data sets are merged using **inner join**
* Data Consolidation is necessary to make a complete sense of the given data.
* Data Understanding
* Combined data set consists of **20,000 rows** and **49 columns**.
* Dataset contains **42 continuous** features and **7 categorical** features.
* Separate EDA for continuous and categorical features is done.
* Majority of the features are continuous (42), so attention to outlier treatment and missing value treatment is given.
* Data Cleaning
  + For continuous features:
    - Outlier treatment at 1and 95 percentile is performed.
    - Outliers are treated at 95% percentile as most continuous features has extreme outliers.
    - For missing value treatment, mean value imputation is used.
* For categorical features:
  + - For missing value treatment, mode value imputation is used.
    - For gender, account\_type and loan\_enq feature, binary encoding (0/1) is used.
    - For income feature, ordinal label encoding (1/2/3) is used.
    - 3 for high ,2 for medium and 1 for low-income level.
    - For region code feature frequency encoding is done.
    - Regions having higher frequency are given higher preference.
* Separating unseen data
  + For **5000** rows, target variable (CC\_cons) is missing, these 5000 rows will act as **unseen data** and prediction for them will be done after final model selection.
  + Rest **15000** records will be used for training, validating and testing of the model.
* Transforming target variable
  + To reduce some skewness of cc\_cons variable, logarithmic transformation is performed.
  + Instead of cc\_cons, log(cc\_cons) will be used in model building.
* Feature Reduction
  + Union of below three feature reduction techniques is done-
    - F-regression
    - Recursive feature elimination (RFE)
    - Select K-best
* Features selected
  + dc\_cons\_jun, cc\_cons\_jun, max\_credit\_amount\_jun, credit\_amount\_jun, investment\_3,

cc\_cons\_apr, age, debit\_amount\_jun, investment\_1, dc\_cons\_apr, credit\_amount\_apr,

Avg\_days\_between\_transaction, cc\_cons\_may, debit\_count\_apr, card\_lim,

max\_credit\_amount\_may, cc\_count\_may, credit\_amount\_may, debit\_amount\_apr,

Emp\_Tenure\_Years, investment\_4, emi\_active, cc\_count\_apr, debit\_amount\_may,

max\_credit\_amount\_apr, dc\_cons\_may, investment\_2

* Model evaluation
  + Root mean square percentage error (RMSPE) is used for model validation
* Model Selection
  + Different types of models used to predict credit card consumption
    - Linear Regression
    - Linear model with regularization -
      1. Lasso or L1 regularization
      2. Ridge or L2 regularization
      3. Elastic Net (L1+L2) regularization
    - Decision tree regressor
    - Ensemble learning-
      1. Bagging regressor
      2. Random forest regressor
      3. Ada boost regressor
      4. GBM regressor
    - KNN regression model
    - SVM regressor
* Model Validation
  + Decile analysis is done for both train and test data
* Final Model selection
  + Support Vector Machine (SVM) regressor model gave the best performance on both train and test data.
  + To predict for unseen data, SVM regressor model is used.