AMATEUR ANALYST

NSSC '19
Data Analytics

PI1926
PI2864
PI2700
PI3140
PI3108

INDEX

INTRODUCTION

Insurance

The insurance industry is an integral part of the global economy. Insurance is a mechanism of risk-transfer from a customer to an insurance company, to protect personal finances in the unfortunate event of accidental loss or damage. It offers full or partial financial compensation in such a case, provided a fee called premium is paid by the customer.
Total permanent disability (TPD) is a condition in which an individual is no longer able to work due to injuries. Total permanent disability, also called a permanent total disability, applies to cases in which the individual may never be able to work again.
Insurance companies classify disability according to the amount of work that an individual is able to perform. Temporary disabilities prevent an individual from working full-time (called temporary partial disability) or at all for a period of time (called temporary total disability). Permanent disabilities prevent an individual from being able to work full-time for the rest of their life, referred to as permanent partial disability, while total permanent disability means that the individual will never work again.
A person will not likely qualify for permanent total disability benefits until the associated medical condition is fixed and stable. What this means is as long as there are additional, curative treatment options available, or a doctor thinks you may improve over time, an insurance company will not call a person “permanently and totally disabled.” Being in this situation doesn't necessarily mean someone won't eventually receive TPD benefits, but it does mean that a person will have to wait until their medical treatment is complete.

Cross-Selling

Cross-sell involves the sale of a new product offered by a single product/service provider to an existing customer. Cross-selling products and services to existing clients is one of the primary methods of generating new revenue for many businesses, including financial advisors. This is perhaps one of the easiest ways to grow their business, as they have already established a relationship with the client and are familiar with their needs and objectives.

OBJECTIVES

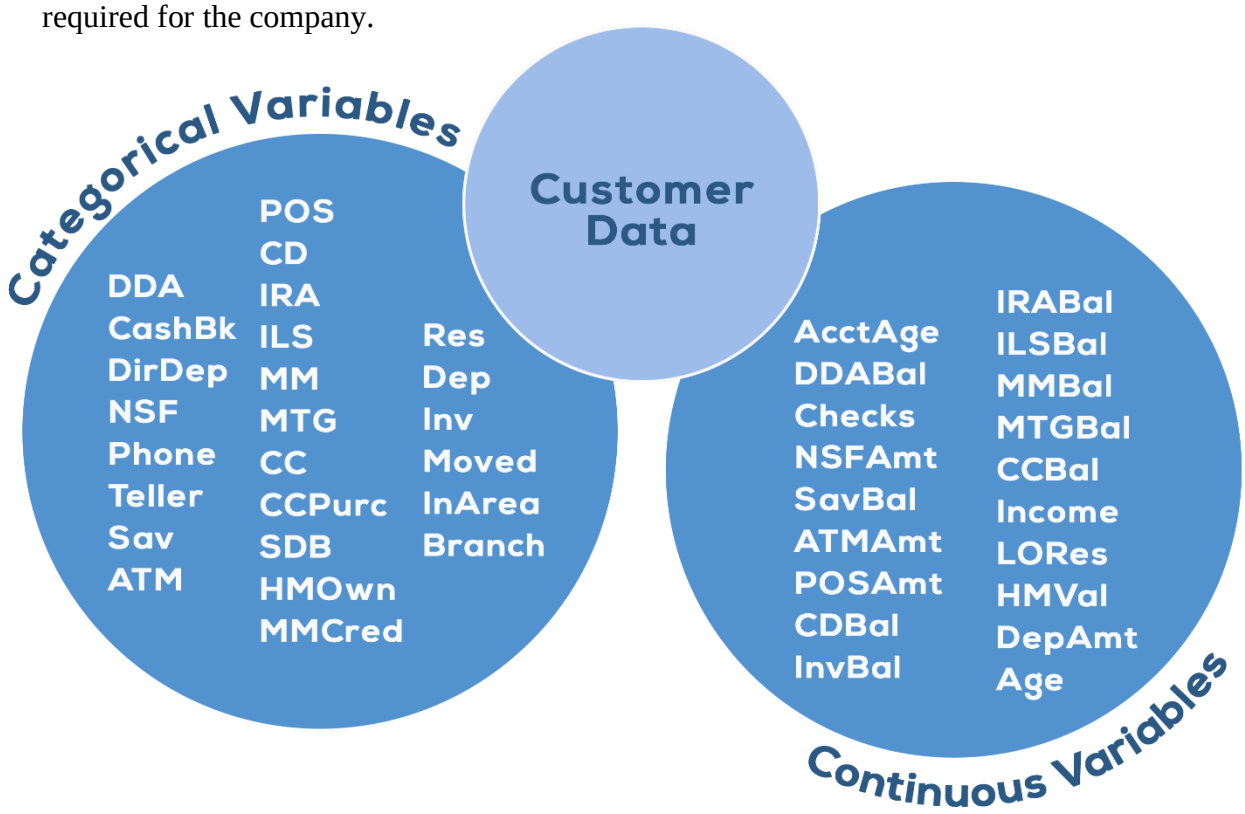
General Objectives - Problem Statement

- To predict the potential customer for their new permanent disability insurance product.
- To provide business strategies for their ideal customer profile

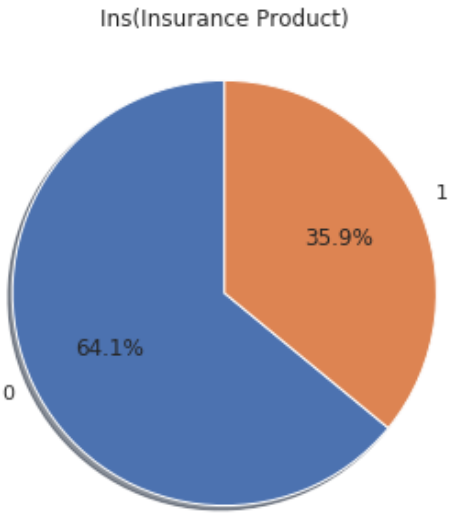
DATA DESCRIPTION

There are 48 variables explaining various characteristics of the customer and the customer-client relationship. The dataset has the potential to gain insight into any of the 32264 customers enlisted. The subjective analysis of the variables present in the financial company statistics gives the entire relationship between customer and the insurance company. The data is provided for predicting whether our customer will buy the new Insurance product or not.

The training data is a pretty balanced dataset containing approximately 65:35 ratio of not potential and potential customers of the product respectively. The user data is perfectly categorized into different columns and covers almost all types of information about customer required for the company.



Target Variable



DATA PREPROCESSING

Data Cleaning

Missing Values: We replaced all the **'.'** in the dataset with NaN hence, had a total of 65125 missing values. Now, handling the missing values can be broadly classified into 2 types which are discussed as follows:

- **Continuous Variables**
  The missing values for variables like Account-Age, POS-Amount, CCBal, CCPurchase, Income, LORes, HMVal, INVbal were replaced by using the mean of the respective columns, excluding the rows which qualify as outliers
  Whereas, for Age, we plotted the difference between the age of the account holder and the age of the former's account and the expected difference came out to be around 43. Hence, NaN in Age was replaced with Account Age + 43

- **Categorical Variables**
  The missing values in variables like POS, CC, HMOwn were replaced in following two ways:
  1. **Using Mode of the categories:** Replaced by the most frequent entry
  2. **Using KNN via Impute:** It replaces the missing values by considering the values of the k- nearest point in a multi-dimensional space.

Outlier Detection: Incorrect or inconsistent data may lead to false predictions and conclusions. That is why before investigating the data, outlier removal needs to be done. We checked for the errors or corruptions or unreal values in the data. But there are no such values in the given data. We tried to detect outliers using **Box-Plots** and **Isolation Forest** and the rows showing significant deviation were removed.

Feature Engineering

NEW FEATURES USING DATA AGGREGATION

a. **Assets:** This is a continuous variable which is calculated by taking into account the various account balances and the amount the customer needs to repay. This variable gives us information about the value of the customer.

   Assets = DDABal + SavBal +CDBal +IRABal + MMBal + Income + HMVal + DepAmt + InvBal -NSFAmt - ILSBal - MTGBal - CCBal

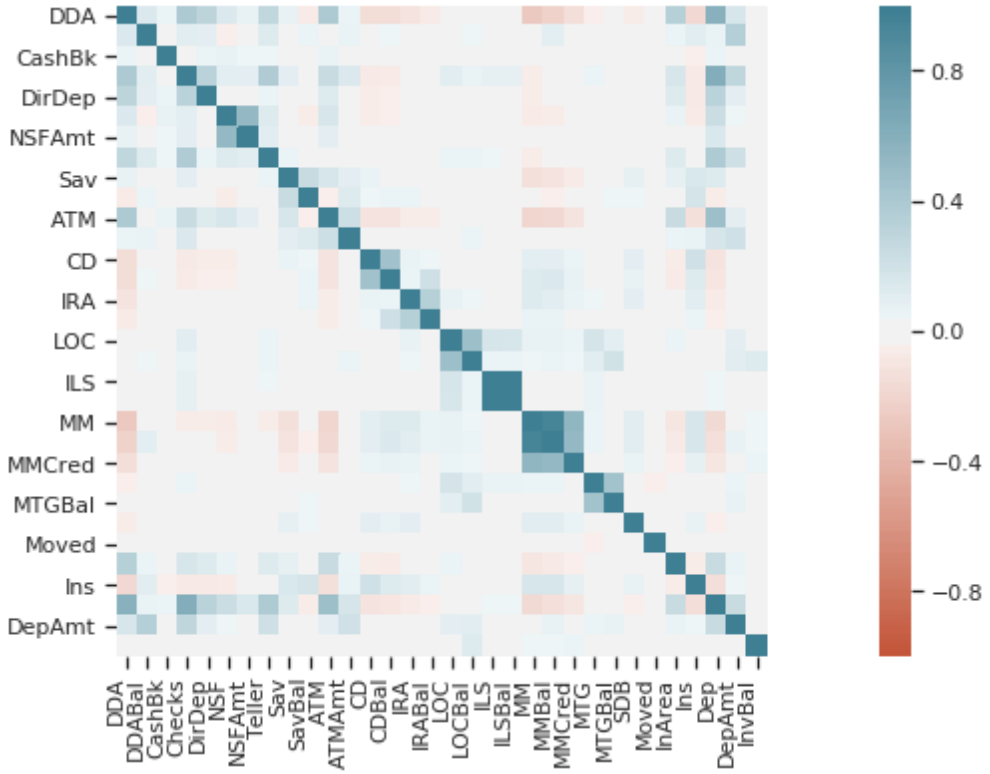
b. **Has Accounts:** This gives the total number of accounts the customer has.

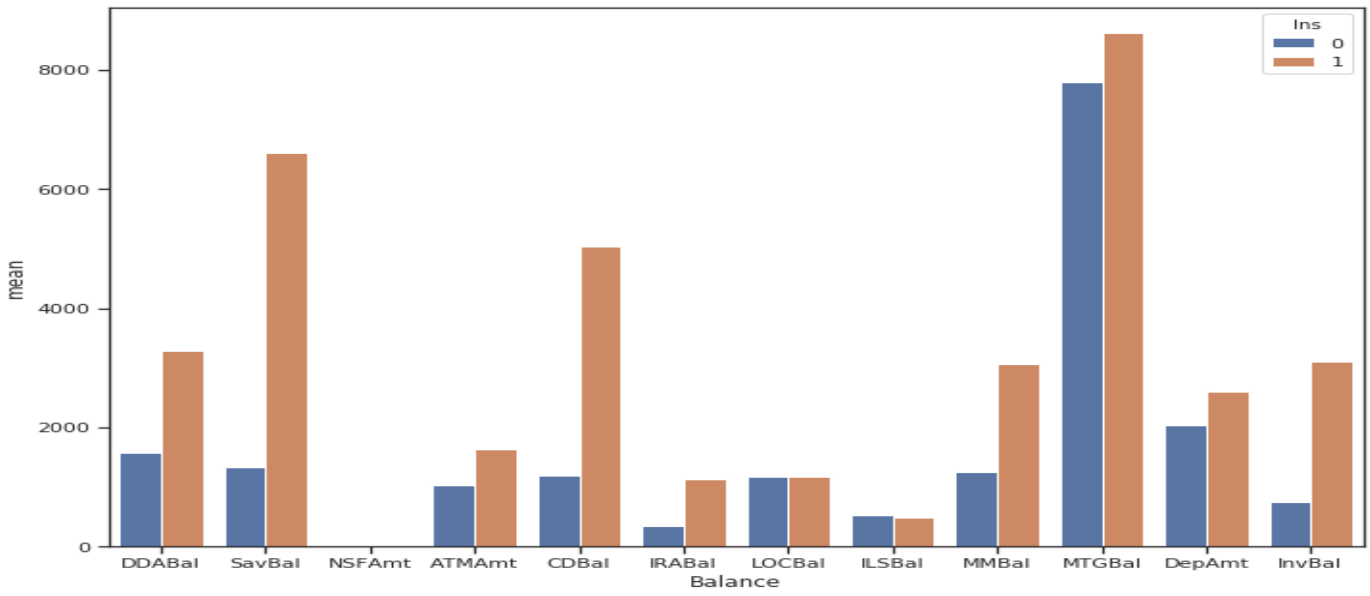
   HasAcc = DDA + Sav + IRA + MM

FEATURE SELECTION

Continuous Variables

Correlation matrix and heat map are used to detect significant features. We also try to find any collinearity amongst the independent variables so as to remove any kind of multicollinearity issue





Following tests were performed to check the significance of the features:

- **Two sample t-test: Alternative hypothesis**
  As you can see columns like LOCBal and ILSBal do not appear to be significant so we do hypothesis testing like t-test or H-test.

  Null Hypothesis (H_0): $\mu_1 - \mu_2 = 0$
  Alternate Hypothesis (H_a): $\mu_1 - \mu_2 \neq 0$

  (where μ_1 and μ_2 are the means of the two samples)

- **Kruskal Wallis test:** Kruskal-Wallis compares the medians of two or more samples to determine if the samples have come from different populations.

  Null hypothesis: Null hypothesis assumes that the samples (groups) are from identical populations.

  Alternative hypothesis: Alternative hypothesis assumes that at least one of the samples (groups) comes from a different population than the others.

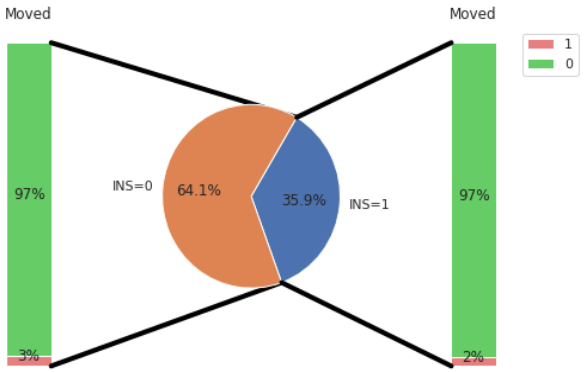
Variables	p-values
LOCBal	0.921
CRScore	0.333
ILSBal	0.053

Categorical Variables

The significant features are selected by first using Data Visualisation Techniques like box plot, histogram and count-plot. After visualizing the graphical representation, we used hypothesis testing to get concrete proof for the same. The hypothesis tests done are:

- **Contingency Chi-Square t-test**
  Null hypothesis: Assumes that there is no association between the two variables.
  Alternative hypothesis: Assumes that there is an association between the two variables.
  If the observed chi-square test statistic is greater than the critical value, the null hypothesis can be rejected.

For feature like Moved the plot appears to be

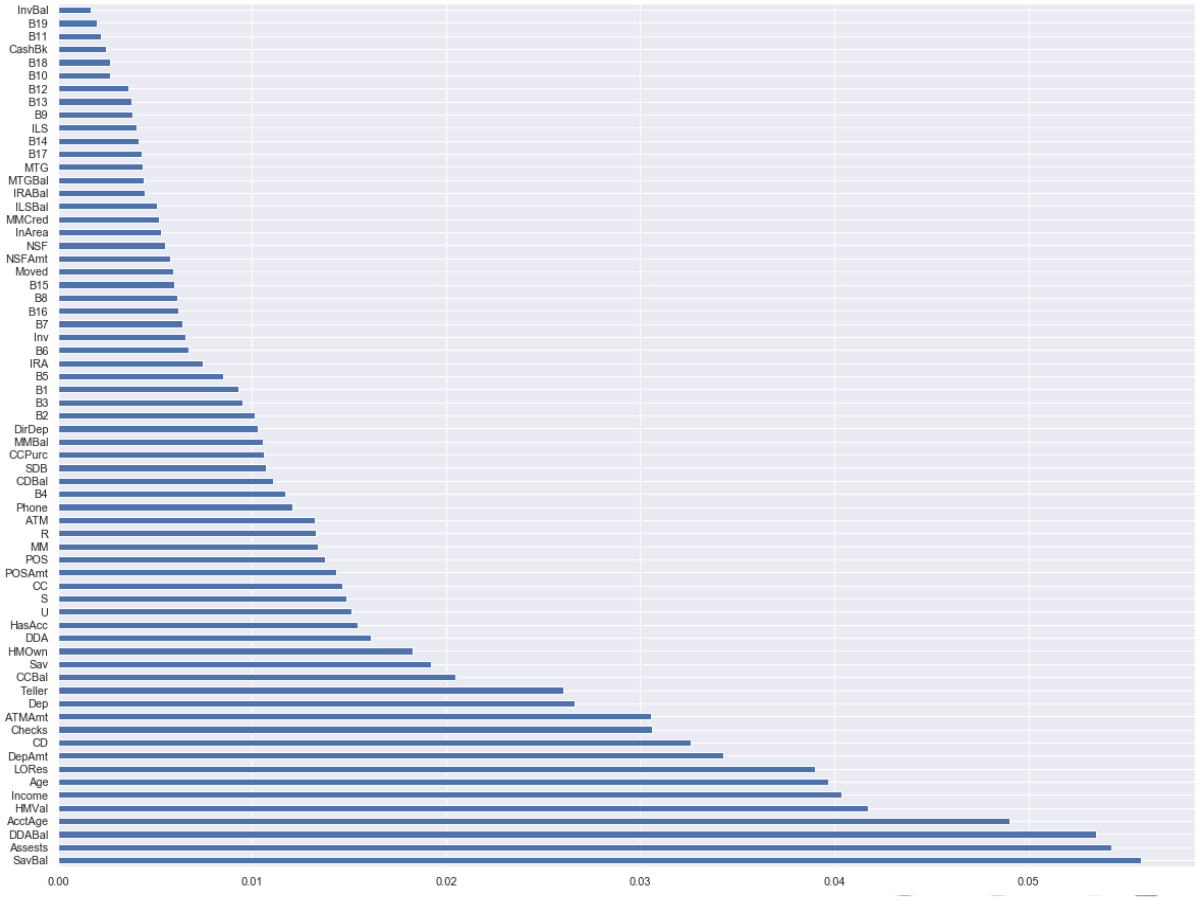


And hence, we use statistical testing to find whether they are significant or not. The p-values of Chi-Square Test for various features are as follows.

Variables	ILS	MTG	CC	CCPurc	HMOwn	Moved
P-Values	0.006	0.961	4.64e-118	7.839e-36	0.922	0.421
Variables	InArea	Branch	Res	Inv	HasAcc	LOC
P-Values	3.232e-26	3.378e-67	0.090	2.2e-59	2.36e-219	0.624

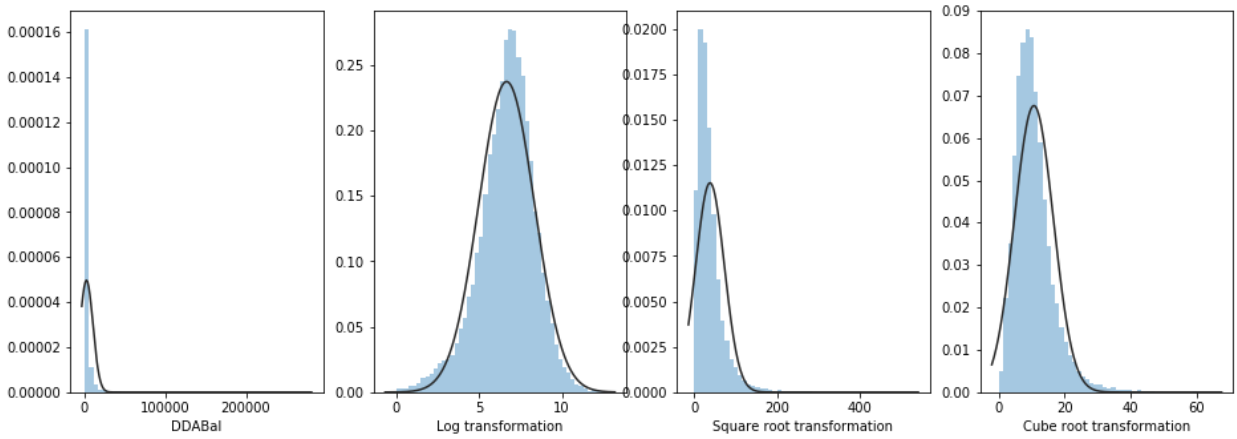
Both categorical and continuous variables

We used Extra trees classifier Model and Mutual Information Classifier (for ranking of the significance of categorical and non-categorical features combined).



FEATURES ANALYSIS

- The features like Checks, DDABal appear to be very skewed. Hence, we use different transformations (Box Cox) and logarithmic, square root, cube root and selected the log as it gives the **most standardized distribution.**



- However, the skewness of some features like AccountAge weren't apparent. Hence, Welch Two Sample t-test was performed on them. And they were standardized based on the p-value.

- The columns Age, CRScore, LORes clearly showed no kind of skewness hence, no transformation was used for the same.

MACHINE LEARNING MODEL

Model Training

Nine different models were tested, namely:

1. Random Forest Classifier
2. Extra trees Classifier
3. Light GBM
4. Gradient Boosting
5. CatBoost
6. Bagging
7. SVC
8. Naive Bayes
9. Logistic Regression

After testing these models, the following models performed decently:

1. **Catboost**
2. **XGBoost Classifier**
3. **Random Forest Classifier**

The pros and cons of the models are as follows:

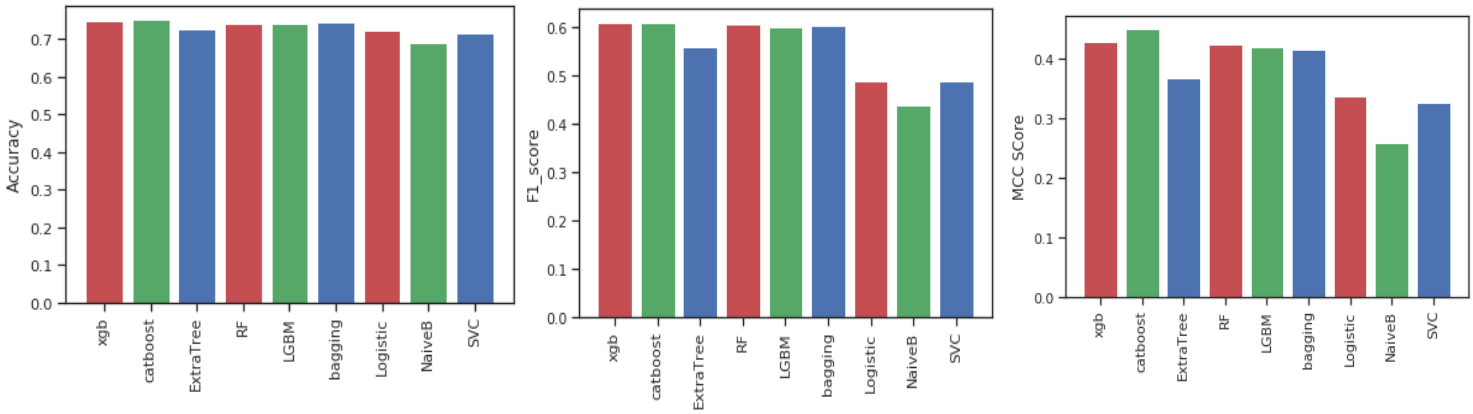
1. **Random Forest Classifier**
   Pros:
   a. It is robust to outliers (which are many in number in our case)
   b. It is also indifferent to non-linear data
   c. Each Decision Tree has high variance, but low bias. But because we average over all trees, we have low bias and moderate variance model
   Cons:
   a. It can tend to overfit and hence needs parameter tuning
2. **XGBoost Classifier:**
   Pros:-
   a. It has the inbuilt capability to handle missing values
   Cons:
   a. Model interpretability: It appears to be a blackbox and is very difficult to be interpreted
3. **Catboost**
   Pros:
   a. CatBoost has the flexibility of giving indices of categorical columns so that it can be encoded as one-hot encoding using one_hot_max_size (Use one-hot encoding for all features with a number of different values less than or equal to the given parameter value).
   b. CatBoost uses an efficient method of encoding which is similar to mean encoding but reduces overfitting.
   Cons:
   CatBoost performs well only when we have categorical variables in the data and we properly tune them and has little or no effect on continuous variables.

We also tried using PCA by selecting the number of clusters on the basis of their explained variance. But that resulted in poorer accuracy in both training and validation set

NOTE: We used both label encoding and one-hot encoding for categorical variables and it was seen that one-hot encoding had better accuracy over both training and validation set.

Results

Model Name	Test Accuracy	F1-Score	MCC (Matthew Correlation Coefficient)
Catboost	0.753	0.610	0.45
XGBoost	0.75	0.609	0.428
Random Forest	0.748	0.608	0.424



Conclusion

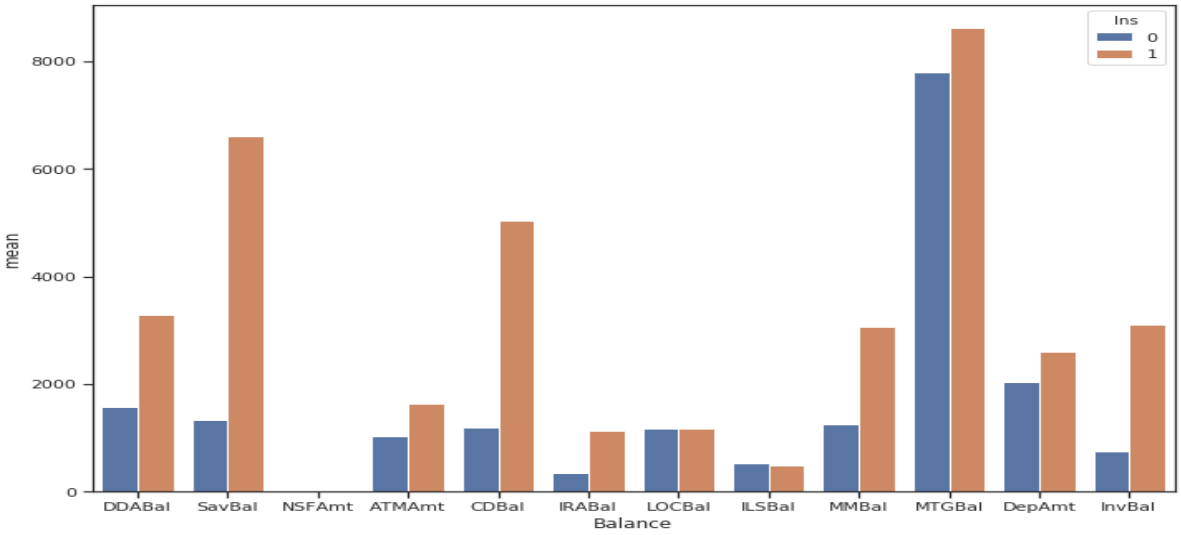
1. Variables like CRScore, LOC, LOCBal and Moved does not effect Insurance product to a great extent, so we ignore these variables in predicting the Insurance Product.

2. We find from Extra trees Classifier that the variables SavBal and Assets are most significant in predicting if the customer is a potential candidate for getting the insurance product.

3. The models which performed well in predicting insurance product are Catboost, XGBoost and Random Forest with Catboost having the highest test accuracy approximately 75.3%.

BUSINESS AND MARKETING INSIGHTS

From our analysis on the data, we find few of the features very important from the business perspective as these are important in determining if the customer is a potential buyer of the insurance product which is a permanent disability product with a low monthly premium. The following features were the most important:



1. **DDABal** - DDABal is the balance a customer has in his checking account. We find that it has a direct impact on the insurance product. The people having more balance(>3500) in their Checking Accounts were in our set of potential customers. Moreover, the marketing team should not focus on people having balance (<1973).

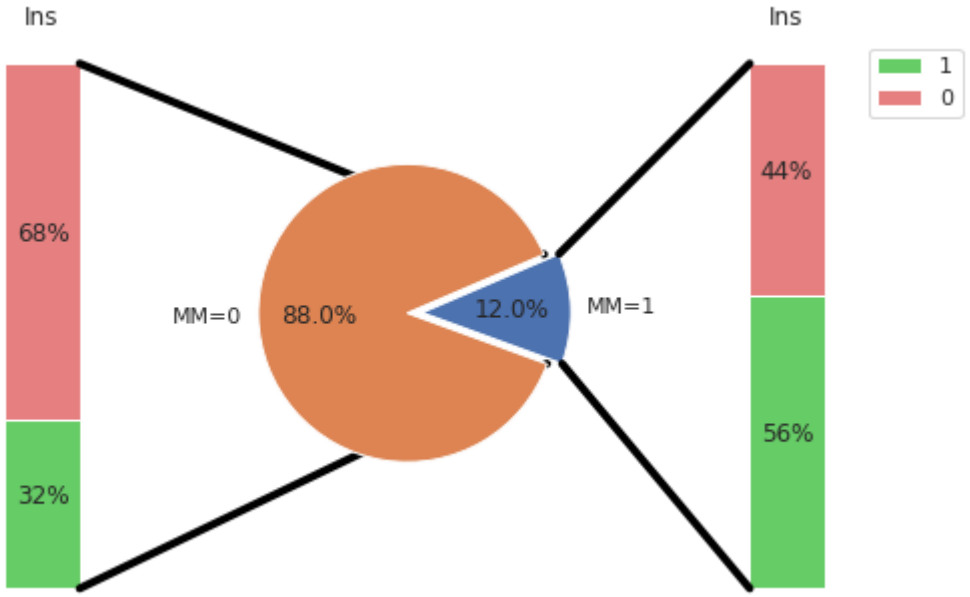
2. **SavBal** (Savings Account Balance) - The classic Savings Account which is very common among masses also appears to be a very significant feature to find the potential customers. The customer who is identified as candidates for the insurance product has a higher savings account balance than the ones that are not

3. **CDBal** (Certificate of Deposit Balance) - A certificate of deposit, or **CD**, is a type of federally insured savings account that has a fixed interest rate and fixed date of withdrawal, known as the maturity date. The balance in these accounts also show interesting results for the comparative study of the clients who are and aren't considered for the cross-selling of the new insurance product

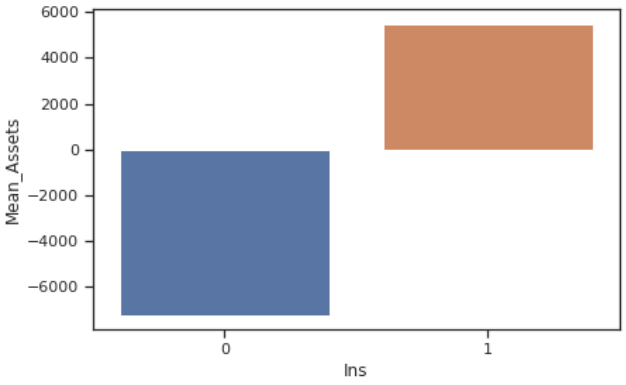
4. **MMBal** (Money Market Balance) - A money market account (MMA) or money market deposit account (MMDA) is a deposit account that pays interest based on current interest rates in the money markets. As seen in the above plot, this also appears to be significant

5. **MM** (Money Market)- The money market is the trade-in short-term debt investments and provides short-term funds. From the plots, we see that people associated with Money Market (MM) have a higher chance of being a potential customer for the

insurance product. Thus, we see that people who are already associated with various investments are more likely to buy the insurance product.

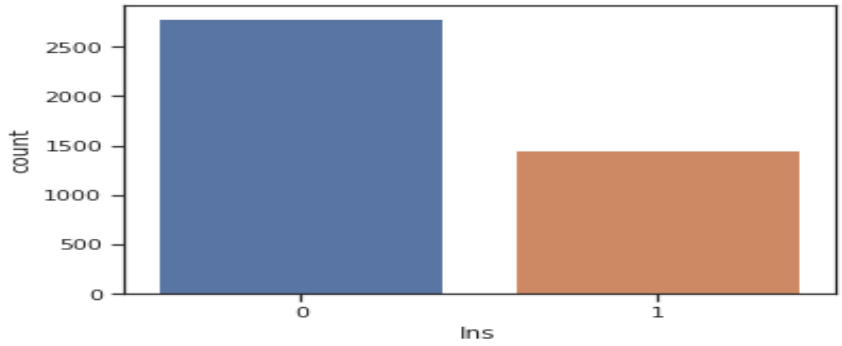


6. **Phones** - Phones is a categorical variable which gives us the number of phones associated with the customer. The greater the number of phones lesser is their chance of being a potential candidate. People having a lot of phones are not preferred for insurance product

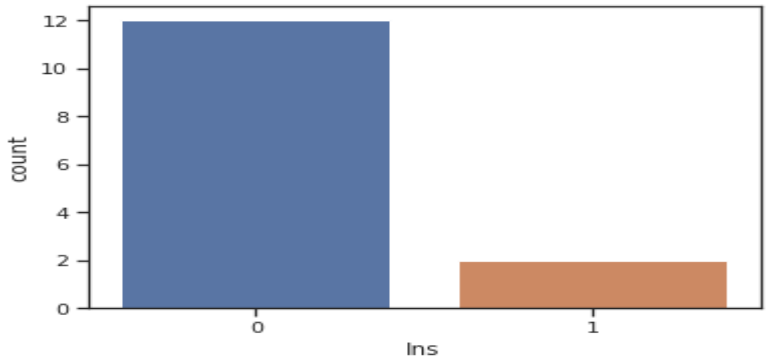


7. **Assets** - The new crested variable also comes out to be an excellent feature to determine whether the client should be entertained for the cross-selling product

8. The customers who are very young and don't own a house are also not considered as potential customers for our new insurance product. Also, the customers who are young and have zero income are also not potential customer.



Above plot shows Ins for Customers having no home that is HMOwn = 0 and age < 45



Above plot shows Ins for Customers having no Income, that is Income = 0 and age < 45

ANNEXURE