Porto Seguro’s Safe Driver Prediction

1. **Objective**

The aim of this problem statement is to predict the probability whether a driver will make an insurance claim, with the purpose of providing a fairer insurance cost on the basis of individual driving habits.It is sponsored by Porto Seguro - a major car and home insurance company in Brazil.

2. **Preparations**

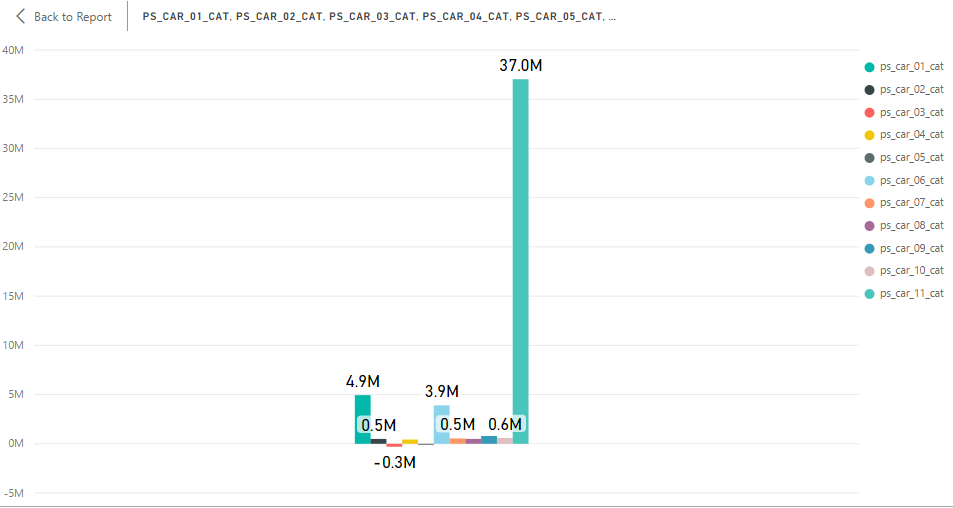
Load the necessary data

3. **Understanding data**

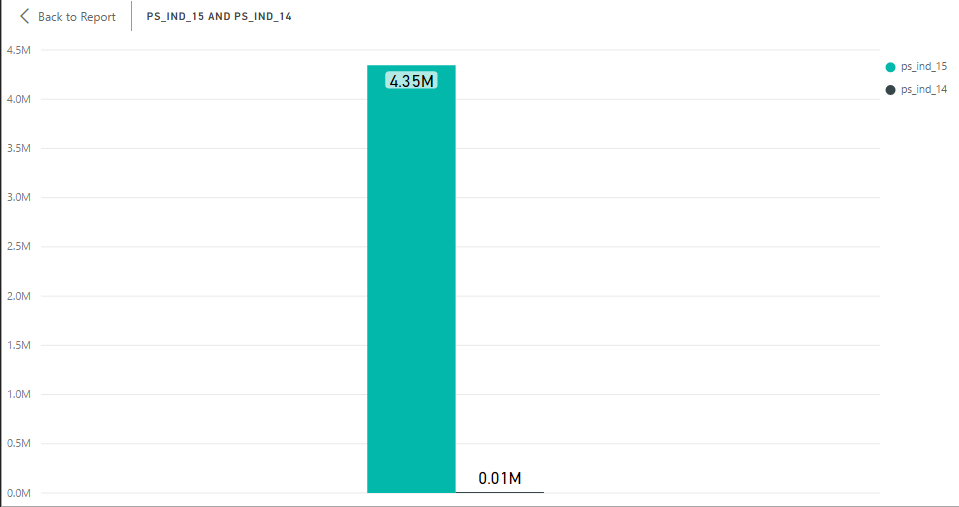
As per the dataset specification, there are binary features, categorical features and numeric features.The data types of the variables are changed according to the specification. Non-numeric categorical value is changed to numeric categorical value. The variable names with- “Ind" is related to individual or driver, “reg” is related to region, “car” is related to car and “calc” is an calculated feature.

**Initial insights**

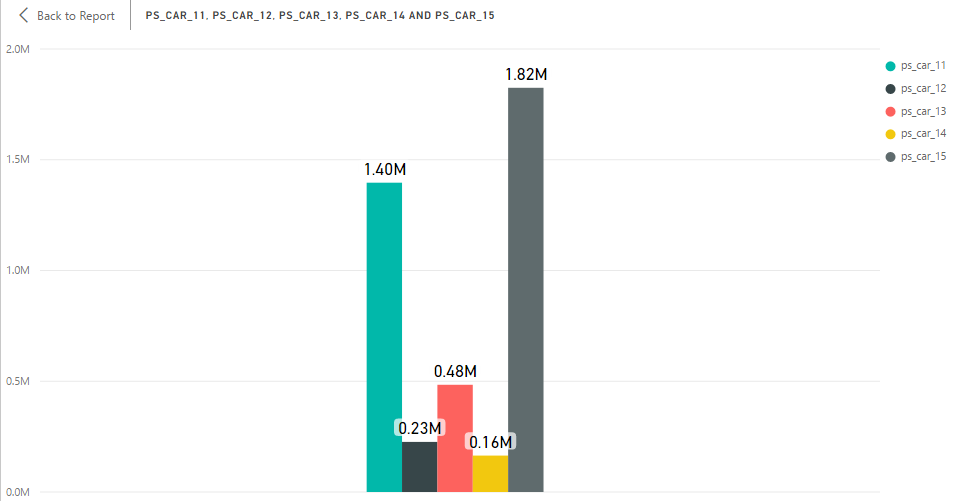
To understand the distribution of features, graphs are plotted for each feature against the target. The graphs of features with related terms are constructed and compared using Power BI, visualization tool.



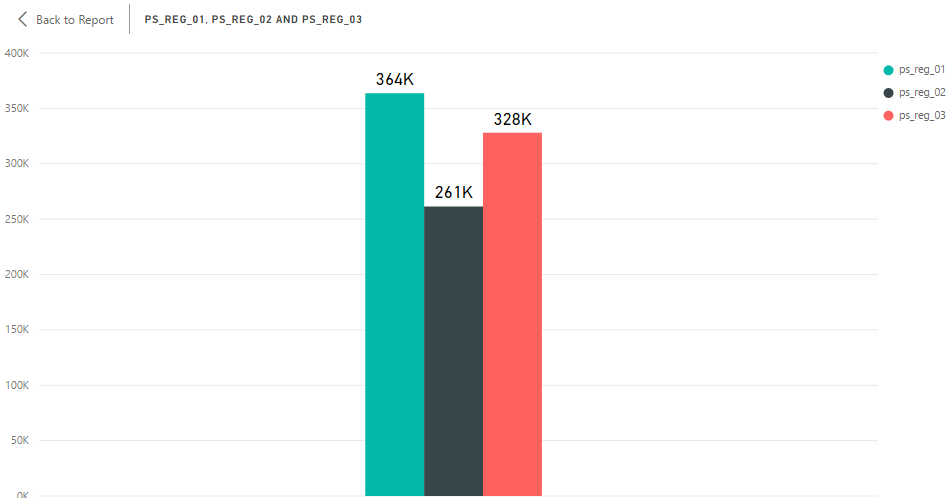
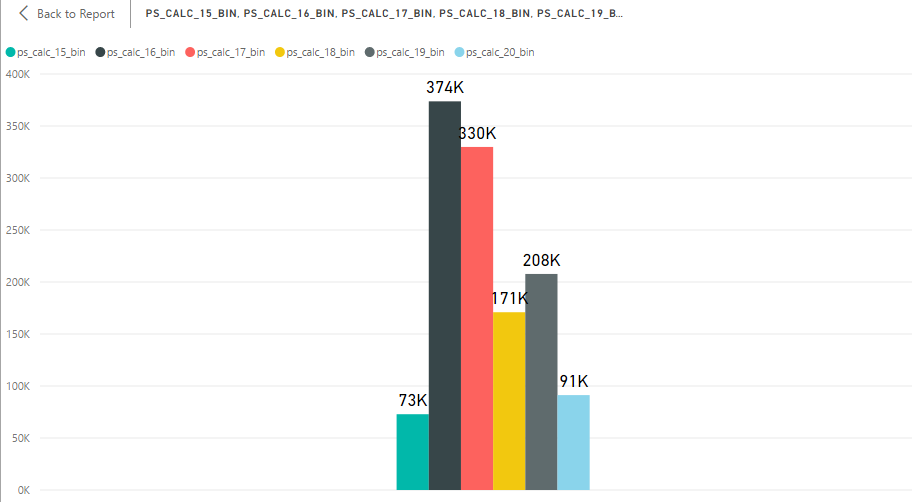
From the above graph, we can find that ps\_car\_06\_cat contributes more information among categorical features of car.

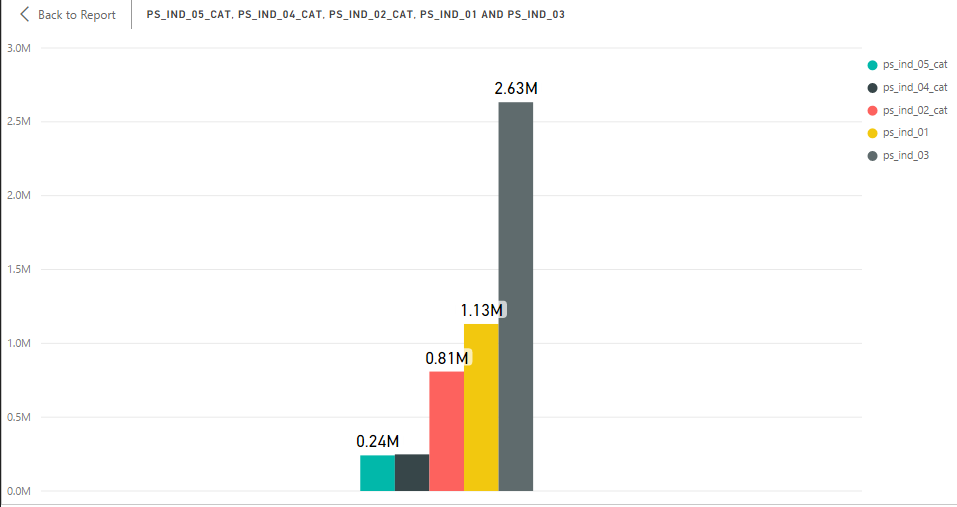


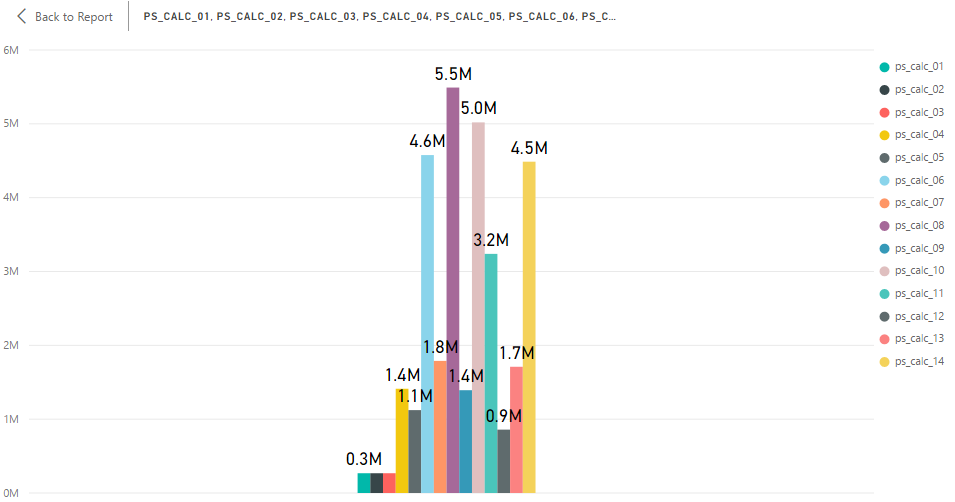
From the above graph, we can find that ps\_IND\_15 contributes more information than the individual feature ps\_IND\_14

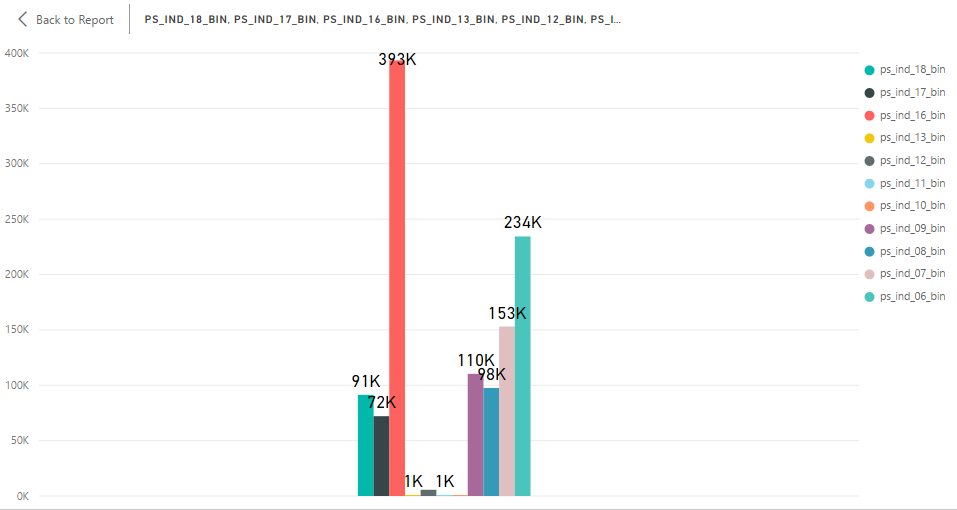


From the above graph, we can find that ps\_car\_15 contributes more information among the group of car related features.

 From the above graph, we can find that ps\_reg\_01, ps\_reg\_02 and ps\_reg\_03 more or less contribute same level of information From the above graph, we can find that ps\_car\_16\_bin and ps\_car\_17\_bin contributes more information among binary features of car.

 From the above graph, we can find that ps\_ind\_15\_cat contributes more information among categorical features of individual.

 From the above graph, we can find that ps\_calc\_06 and ps\_calc\_10 contributes more information among calculated features of car.



From the above graph, we can find that ps\_ind\_16\_bin contributes the highest level of information among the individual features which are in binary.

**Missing values**

1. Missing values are analyzed column wise. ps\_car\_03\_cat contains 411231(almost 90%) missing values, so this column is removed. ps\_car\_05\_cat contains 266551 (44%) missing values, so this column is also removed.
2. Imputing missing values with Median and Mode as per the data type of the columns.
3. The rows with negligible missing values are removed.

**4. Outlier analysis**

Outlier analysis is done to remove the outliers by standardization and normalization process.

**5**. **Data subset**

Due to memory limitations of the system a major subset of the data is taken for further analysis.

**6. Correlation analysis**

On running the correlation analysis for numerical variables, it is observed that most features appear to be primarily correlated with others in their group. We could find that, there is a strong correlation between the variable ps\_ind\_12\_bin and ps\_ind\_14 with 0.94. And correlation values with above 0.55 are between features ps\_reg\_03 & ps\_reg\_02, ps\_car\_12 & ps\_car\_13,ps\_car\_14 & ps\_car\_12, ps\_car\_15 & ps\_car\_13. Also, there is a negative correlation between ps\_ind\_16\_bin and ps\_ind\_18\_bin with -0.58.

**7**. **Chi-square test**

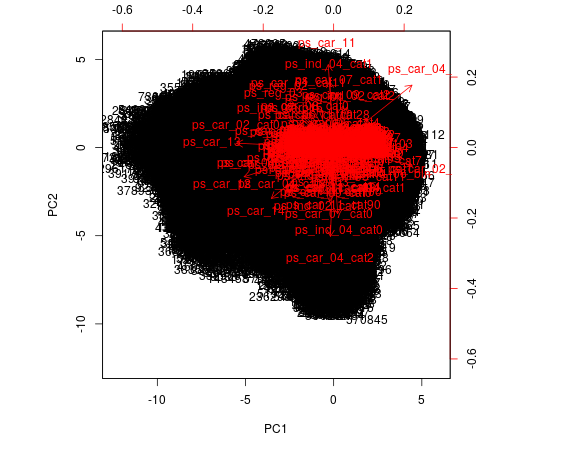
For categorical values, chi-square test is carried out to understand the significance of the variables.

**8**. **Feature Engineering**

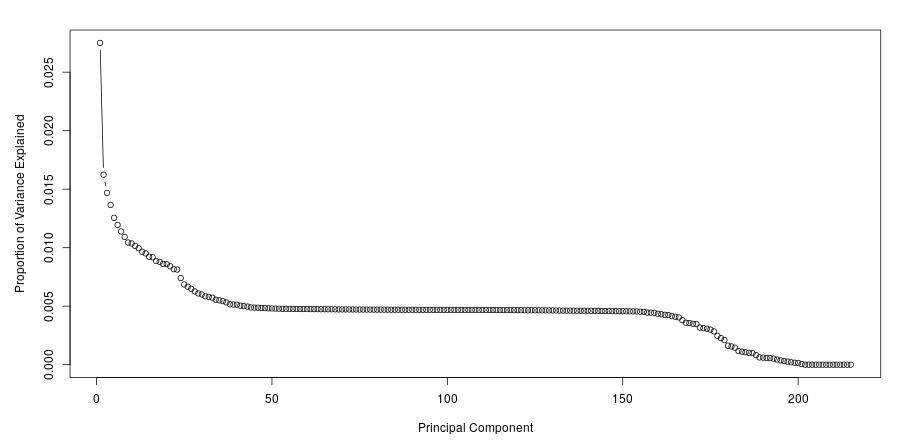
Principal Component Analysis:

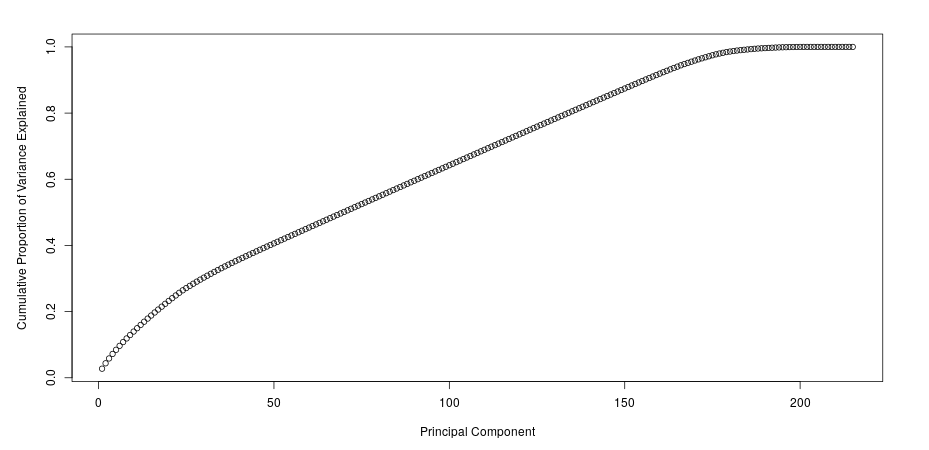
Principal component analysis is used to extract important variables from high dimensional data. The matrix on which it is performed should be numeric. So, the categorical variables are converted to numerical by creating dummy data. Principal components give components like PC1, PC2, etc. depicts the linear combination of features. PC1 captures the variables that describe high variance; PC2 captures the next level of variance in the remaining dataset. The directions of the principal components describe the correlation between the principal components. If PC1 and PC2 are orthogonal, it means they are not correlated. One the principal components are found, model is built. Here, Logistic regression model is used.

PCA is performed on the train data and the resultant principal components are plotted.



To understand the variance level with each principal component, scree plot and cumulative scree plot is used.





**9. Train and Test data**

Using stratified sampling the dataset is split into train and test data at this stage.

**10. Model Building**

PCA is constructed for test data.

Logistic regression model: Logistic regression model is used in this problem statement, to calculate the probability of claiming insurance. This model is preferred because the target variable is categorical in nature. And also to predict the probability of occurrence of claiming insurance by fitting data to a logit function.

**Process:** It’s a part of class of algorithms known as Generalized Linear Model. Here the generalized linear model equation is –

g(y) = Bo + B1X1 + B2X2 +.....

The link function is established using probability of success (p) and the probability of failure (1-p).

i.e. Link function y= log (p/ (1-p))

From the above two equations, the logistic regression equation is

Log (p/1-p) = Bo+B1X1

**Implementation:**

R uses glm() function where the parameter link = logit invokes logistic regression model on the given set of data.

**11**. **Model Evaluation**

The predicted result is the probability of occurrence of target variable.

Optimal Cutoff value is calculated with respect to the target variable and predicted variable. The variable is found to be 0.25.

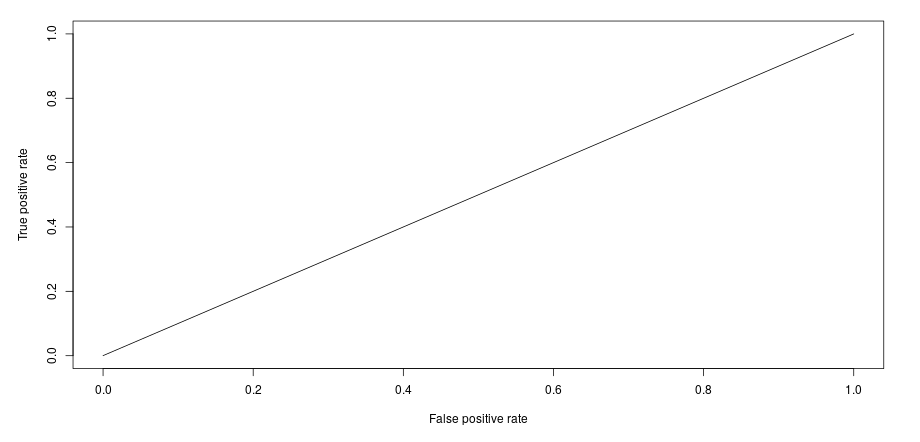
With respect to this optimal cut off value, confusion matrix is built.

Specificity is found to be - 0.9999942

Sensitivity is found to be - 0.000153657

Model accuracy is found to be close to 96%

ROC is fit and the Area under the curve is calculated.



**12**. **Business Insights**

From the model building we can infer that the probability of not claiming the insurance is high. Meaning, the company has insurers who are good drivers. So, to gain more profit the insurance company can increase the cost of insurance to gain more profit.

By

Priyadharsini