**Safedriver prediction comparison using various methods**

**Dataset Description:**

* **Dataset Name:** Insurance.csv(train.csv from kaggle safedriver prediction)

The dataset is about the largest auto and homeowner insurance companies, [Porto Seguro](https://www.portoseguro.com.br/) in Brazil. Inaccuracies in car insurance company’s claim predictions raise the cost of insurance for good drivers and reduce the price for bad ones. It contains the data, where each row corresponds to a policy holder, and the target columns signifies that a claim was filed.

I am to build a model that predicts the probability that a driver will initiate an auto insurance claim in the next year. A more accurate prediction will allow them to further tailor their prices, and hopefully make auto insurance coverage more accessible to more drivers.

* **Columns and Rows:** 595k rows and 59 columns

**Problem Statement:**

The aim of the study is to predict the probability whether the driver will make an insurance claim with the purpose of providing the fairer insurance cost on the basis of individual driving habits. I will be predicting whether a claim will be filed (target = 1) or not (target = 0). The dataset comes under binary classification and dataset being imbalanced is subjected to balanced condition by applying various features. So I am to set up models according to the dataset and predict whether the drivers will initiate an insurance claim in the next year.

**Table Columns are:**

Id, Target, ps\_ind\_01, ps\_ind\_02\_cat,ps\_ind\_03,ps\_ind\_04\_cat,ps\_ind\_05\_cat,

ps\_ind\_06\_bin, ps\_ind\_07\_bin,ps\_ind\_08\_bin,ps\_ind\_09\_bin,ps\_ind\_10\_bin,

ps\_ind\_11\_bin, ps\_ind\_12\_bin,ps\_ind\_13\_bin,ps\_ind\_14,ps\_ind\_15,

ps\_ind\_16\_bin, ps\_ind\_17\_bin,ps\_ind\_18\_bin,ps\_reg\_01,ps\_reg\_02,

ps\_reg\_03ps\_car\_01\_cat, ps\_car\_02\_cat, ps\_car\_03\_cat,ps\_car\_04\_cat,

ps\_car\_05\_cat, ps\_car\_06\_cat, ps\_car\_07\_cat, ps\_car\_08\_cat, ps\_car\_09\_cat,

ps\_car\_10\_cat, ps\_car\_11\_cat,ps\_car\_11,ps\_car\_12,ps\_car\_13,ps\_car\_14,

ps\_car\_15, ps\_calc\_01, ps\_calc\_02, ps\_calc\_03, ps\_calc\_04,ps\_calc\_05,

ps\_calc\_06, ps\_calc\_07, ps\_calc\_08, ps\_calc\_09, ps\_calc\_10, ps\_calc\_11,

ps\_calc\_12, ps\_calc\_13, ps\_calc\_14, ps\_calc\_15\_bin, ps\_calc\_16\_bin,

ps\_calc\_17\_bin,ps\_calc\_18\_bin, ps\_calc\_19\_bin, ps\_calc\_20\_bin.

**Column Details:**

In the table, the features that belong to similar groupings are tagged as such in the feature names (e.g. ind,reg,car,calc). In addition, feature names include the postfix **bin** to indicate binary features and **cat** to indicate categorical features. Features without these designations are either continuous or ordinal. Values of -1 indicate that the feature was missing from the observation. The target column signifies whether or not a claim was filed for that policy holder.

**Result:**

I have studied and analyzed the dataset and applied various plotting techniques and features for the prediction. Setting up five models with respect to the dataset availed for study. Among such models the model with respect to logistic regression was considered as the suitable model for our prediction.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **F1 Score (%)** | **Accuracy Score (%)** | **Recall Score (%)** | **Precision Score (%)** |
| **Logistic Regression** | 56.74 | 58.60 | 54.84 | 58.77 |
| **Random Forest** | 49.54 | 54.78 | 44.84 | 55.35 |
| **Decision Tree** | 51.18 | 55.19 | 47.45 | 55.55 |
| **KNN** | 53.21 | 54.16 | 52.64 | 53.79 |
| **SVC** | 54.91 | 58.75 | 50.72 | 59.85 |

**Conclusion:**

The journey of project was having some sort of mysterious nature. The dataset availed was an unbalanced one containing 595k rows and 59 columns to predict the probability whether the driver will make an insurance claim with the purpose of providing the fairer insurance cost subjected to the company. In the preprocessing stage i have dropped less correlated columns, the columns having null values and also replaced the null values with NaN values. The most important step regarding my study was random under sampling because the dataset was highly imbalanced. However i tackled it taking equal set of zero value rows and one target value rows by doing under sampling .Thus regarding my study by choosing logistic regression model as the suitable one, i came to the conclusion that more drivers will initiate an auto insurance claim in the next year than the current year.