C22070780

CMT-307 Applied Machine Learning
Implementation and Evaluation of a Case Study Using Machine Learning Techniques
Student Number-C22070780

INTRODUCTION

A health insurance provider is looking to expand into car insurance. We have been given a dataset in which our task is to build Machine Learning Models to predict the Result of whether a customer will buy Car Insurance based on the given variables.Based on the information from the column names and associated data values, the dataset contains categorical and numerical features as follows.

Categorical Features

• Gender- Gender of the customer

• HasDrivingLicense - If the customer has a driving license or not

• RegionID- Unique code for the region customer is from

• Switch- If the customer had a previous insurance

• VehicleAge- The age of the vehicle

• PastAccident- If the vehicle was in an accident in the past or not

• SalesChannelID- This is a numeric column

Numerical Features

• Age- Age of the customer

• AnnualPremium- The amount the customer needs to pay in a year

• DaysSinceCreated - Number of days the customer has been associated with the company.

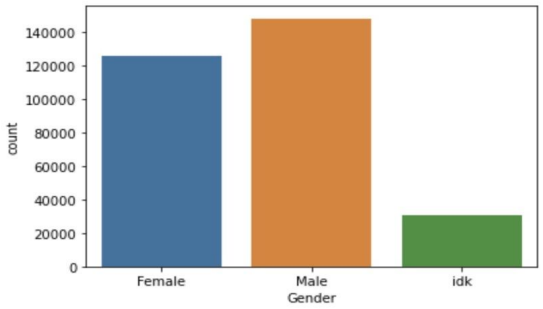
EXPLORATORY DATA ANALYSIS

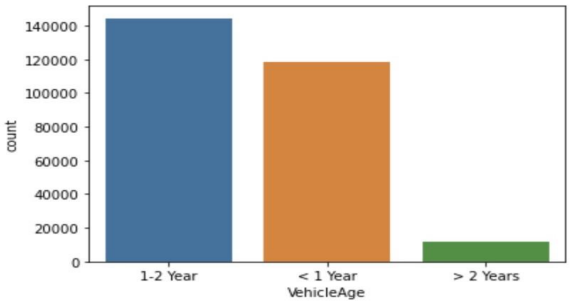
As the preliminary step, we import all the required libraries to carry out the project and start out basic analyses of the data set. This step gives us a clear idea of the kind of dataset we are working with. After importing all the required modules, the first few commands are to find out the basic insights regarding the dataset viz., shape, size, description of the data frame, data types of the columns used, dropping Null/NA values etc. To check which feature is correlated with the target attribute we plot a heatmap of every feature with respect to the Result(Target Feature)

We plot different graphs like boxplot,scatterplot,displot and catplot with various attributes to understand the data and carry out the various Preprocessing Analysis.

C22070780

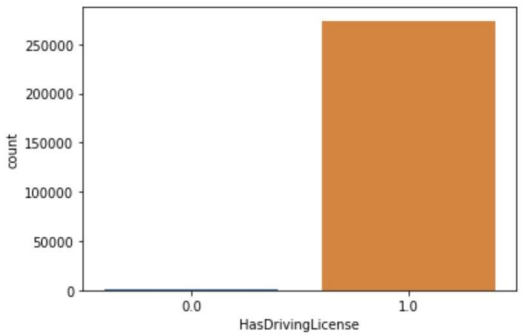
From the various plots it can be implied that Annual Premium has a positively skewed distribution, we can also depict that DaysSinceCreated has an approximately uniform distribution.Age Column has some outliers but we are not going to treat them as it would not be affecting our results.

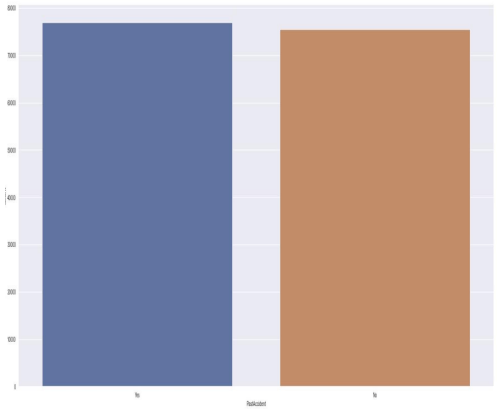




Correlation of result with other attributes:

```
Result                1.000000
PastAccident          0.108470
Age                   0.050515
Gender                0.037671
AnnualPremium         0.022547
RegionID              0.007823
HasDrivingLicense     0.000369
DaysSinceCreated     -0.001135
id                   -0.002333
VehicleAge           -0.060519
Switch               -0.109460
SalesChannelID       -0.137757
Name: Result, dtype: float64
```





C22070780

DATA PRE-PROCESSING

In this step, we clean the data and either remove missing values or find the mean value for the column using the fillna() method. Then the code uses isna().sum() method to sum the number of null values in each column and returns the result.

We convert all the data types into numeric so it is easy for data implementation. Label Encoding is the process used to carry out the transformation of Object type attributes into int/float type.

For Annual Premium, we drop the £ symbol using the lambda function. We use Standard Scaler in Preprocessing.

For Gender we divide into three categories and find the mean for the unknown data and fill the unknown values.

For region, Switch, VehicleAge, PastAccident and HasDrivingLicense we fill the unknown values.

We drop rows which have more than or equal to 4 NA values to remove the noisy data and clean the dataset.

MODEL IMPLEMENTATION

We combine all the feature names with which we are going to build the model.

After the transformation of the required features, they are split into Testing and Training data x_train, y_train, x_test and y_test.

Imbalanced classification involves developing predictive models on classification datasets that have a severe class imbalance. One approach to addressing imbalanced datasets is to oversample the minority. For the same, a technique known as SMOTE (Synthetic Minority Oversampling Technique) has been used.

For Model Implemenation we have used three classification algorithms:

1. Random Forest Classifier
2. K Nearest Neighbors
3. Decision Tree Classifier

1. Random Forest Classifier has been used because - Accuracy of Random forest is generally very high, provides an estimate of important variables in classification, forests generated can be saved and reused, unlike other models It does not overfit with more features.
2. *k*-NN is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy dramatically. It's ideal for non-linear data since there's no assumption about underlying data. It can naturally handle multi-class cases. It can perform well with enough representative data.
3. Decision Tree Classifier is used for the following reasons - Their outputs are simple to read and analyse, and they don't require statistical expertise. Decision trees need fewer data preparation

efforts than other decision procedures. Users must have available information in order to build new variables with the ability to forecast the target variable; however, once the variables have been formed, fewer data cleaning is necessary. Missing values and outliers have less impact on the data in the decision tree.

PERFORMANCE EVALUATION

The performance of the model is evaluated using Accuracy, Recall, F1 Score, and Precision as the metrics. The ROC_AUC score is also taken into account.

Random Forest Classifier

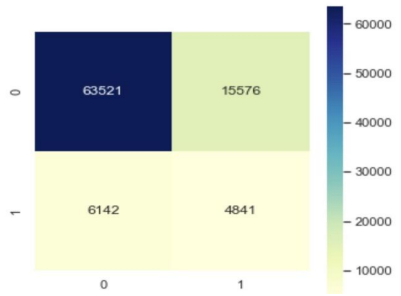
```
ACCURACY OF THE MODEL:  0.769971136767318

              precision    recall  f1-score   support

           0       0.92      0.81      0.86     79097
           1       0.26      0.48      0.34     10983

    accuracy                           0.77     90080
   macro avg       0.59      0.64      0.60     90080
weighted avg       0.84      0.77      0.80     90080
```

Confusion matrix for Decision Tree Classifier



The ROC_AUC score is 0.64

K-Nearest Neighbors

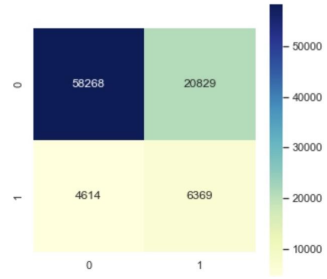
```
ACCURACY OF THE MODEL:  0.7175510657193606

              precision    recall  f1-score   support

           0       0.93      0.74      0.82     79097
           1       0.23      0.58      0.33     10983

    accuracy                           0.72     90080
   macro avg       0.58      0.66      0.58     90080
weighted avg       0.84      0.72      0.76     90080
```

Confusion matrix for K Neighbors Classifier



The ROC_AUC score is 0.65

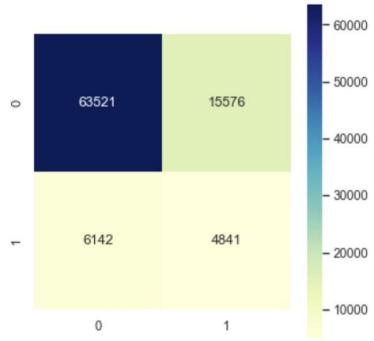
Decision Tree Classifier

```
ACCURACY OF THE MODEL:  0.7589031971580817
              precision    recall  f1-score   support

           0       0.91      0.80      0.85     79097
           1       0.24      0.44      0.31     10983

    accuracy                           0.76     90080
   macro avg       0.57      0.62      0.58     90080
weighted avg       0.83      0.76      0.79     90080

Confusion matrix for Decision Tree Classifier
```



The ROC_AUC score is 0.62

RESULTS AND DISCUSSION

First, we load our dataset and check for null values. There were a lot of null values, so treatment of missing values was done and after data processing, we applied feature scaling techniques to normalize our data to bring all features on the same scale and make it easier to process by ML algorithms. Through Exploratory Data Analysis, we categorized Age, Region_Code, etc. Further, we observed that customers belonging to young Ages are more interested in vehicle response. We observed that customers having vehicles older than 2 years are more likely to be interested in vehicle insurance. Similarly, customers having damaged vehicles are more likely to be interested in vehicle insurance.

The best 3 models were KNN, Random Forest and Decision Tree. We select KNN as our best model considering precision and recall as we have an unequal number of observations in each class in our dataset, so accuracy alone can be misleading. It has a better ROC_AUC score of 0.65.

We were asked to predict if the customer is willing to buy the car insurance, seeing the metrics, the above ML algorithm will likely be more efficient in predicting whether the customer is not willing to purchase the car insurance, because of the highly imbalanced data set or maybe huge missing values.