CMT307 – APPLIED MACHINE LEARNING

C22031444

INTRODUCTION

Machine learning is a branch of computer science that arose from the study of data pattern recognition as well as artificial intelligence's computational learning theory. It's a first-class ticket to the most exciting data analytics jobs available today. As the number of data sources grows, so does the computational power required to process them. Going straight to the data is one of the simplest ways to quickly acquire insights and make predictions. Decision making, clustering, classification, forecasting, deep learning, inductive logic programming, support vector machines, reinforcement learning, similarity and metric learning, genetic algorithms, sparse dictionary learning, and so on areall sub-problems of machine learning. The machine learning task of inferring a function from data is known as supervised learning or classification.

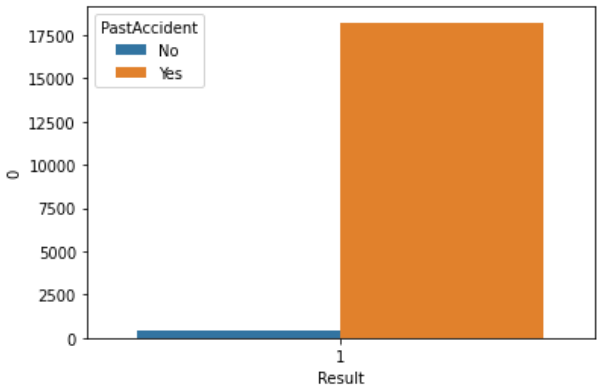
The goal of an Health insurance company to switch to car insurance company which we will examine in this study. The entiredataset is provided, and the end outcome is whether or not the customer will switch to car insurance. This is a classification type of supervised machine learning project.

The following are the steps involved in the topic we'll be discussing:

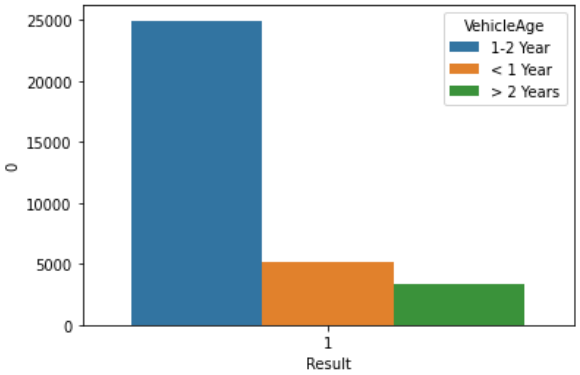
1. Exploratory Data Analysis.
2. Data Pre-processing.
3. Model Implementation.
4. Performance Evaluation.
5. Results and Discussions.

EXPLORATORY DATA ANALYSIS

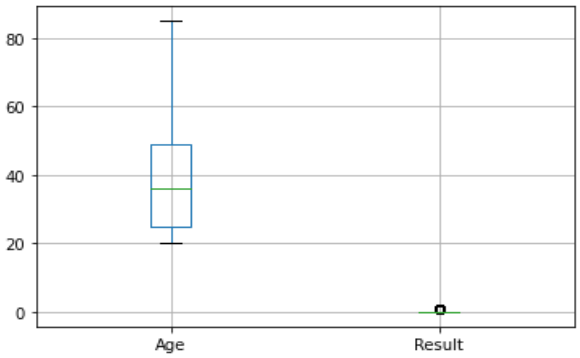
The dataset has the following column names, which are largely self-explanatory. The dataset consists of 304887 rows and 12 features with different types of data types, there are 3 numerical and 7 categorical features and RESULT is our target. There are no duplicates in the given data set.



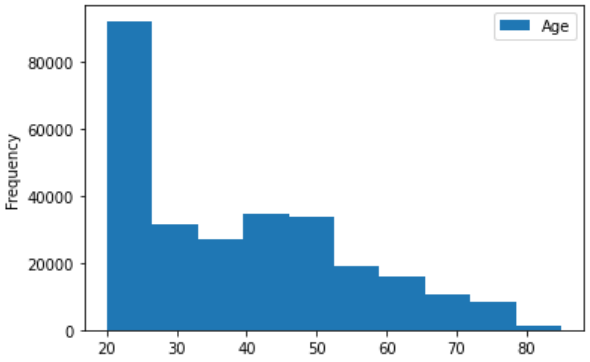
The above barchart represents the people who has past accident history are showing more interest towards buying vehicle insurance.



The above bar chart represents customers who recently bought the vehicle are showing interest in buying vehicle insurance.



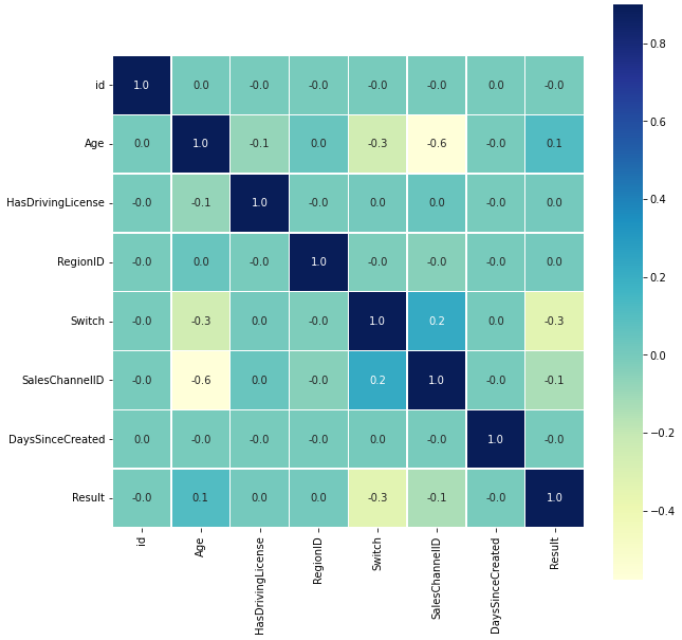
The above Box plot shows there are no outliners.



The above histogram represents population distribution is from age 20 to 40.

Correlation matrix:

It illustrates the outcome is unconnected to any other feature in a positive sense. The feature Switch and result are inversely associated, and HasDrivingLicense strongly connect with Sales ChannelID, showing that SalesChannelID receives queries from clients of all the ages.



Target Feature:

Depending on these factors determine whether a customer will purchase car insurance. The dataset needs to be more balanced, as evidence by the number of people who are not interested (2,67,700) And people who are interested in buying insurance are (37,187)

Statistical Analysis:

- Male customers own more automobiles than female clients. The vehicles are 1-2 years old have been in accidents.
- **Age :** 75 percent of the customers is 49 or younger, with an average of 39 and coefficient of 1, indicating a relatively low variation.
- **HasDrivingLicense :** There are approximately 75 percent of clients having a license.
- **AnnualPremium :** The average yearly premium for a customer who is 39 years old is £1,528.59 and the standard deviation is enormous with a cv > 0.5.
- **RegionID :** There are about 80k people wit health insurance in a region with ID 28 than anywhere else.
- **VehicleAge :** A few thousand cars are older than 2 years.
- **SalesChannelID :** 1,66,897 buyers have inquired through the sales channel ID's 152 and 124.

Missing Values :

- Columns which have missing values in Numerical features are Age, Days Since Created, RegionID and SalesChannelID.
- Cloumns which have missing values ion categorical features are Gender, VehicleAge, PastAccident and Switch.

DATA PRE-PROCESSING

Splitting the data into train and test :

Data is splitted into train and test size = 30% where 70% data is trained to train and 30% data is trained to test.

Null Values Imputation :

- **Mode Imputation :** Gender, VehicleAge, PastAccident, Switch.
- **Mean Impuation :** Age, DaysSinceCreated, RegionID, SalesChannelID.

Feature Scaling :

- For categorical features we have encoded pipeline as OneHotEncoder.
- For numeric features we have encoded pipeline as StandardScaler.

Finally, fit the pipeline object on train data and transform both the train and test data.

MODEL IMPLEMENTATION

For model implementation we have used three classification algorithms – Logistic Regression, KNeighbors Classifier and SVM Classification.

- **Logistic Regression** has been used because – it is one of the simplest machine learning algorithms and is easy to implement yet provides great training efficiency in some cases. Also due to these reasons, training a model with this algorithm doesn't require high computation power. This algorithm allows models to be updated easily to reflect new data. In a low dimensional dataset having a sufficient number of training examples, logistic regression is less prone to over-fitting.
- **K-Neighbors Classifier** has been used because -- easy to apply classification method which implements the k neighbors queries to classify data. It is an instant-based and non-parametric learning method. In this method, the classifier learns from the instances in the training dataset and classifies new input by using the previously measured scores.
- **RandomForest Classifier** has been used because- It can perform both regression and classification tasks. A random forest produces good predictions that can be understood easily. It can handle large datasets efficiently. The random forest algorithm provides a higher level of accuracy in predicting outcomes over the decision tree algorithm.

PERFORMANCE EVALUATION

Performance of the model is evaluated using Accuracy, Recall, F1 Score, and Precision as the metrics

Logistic Regression

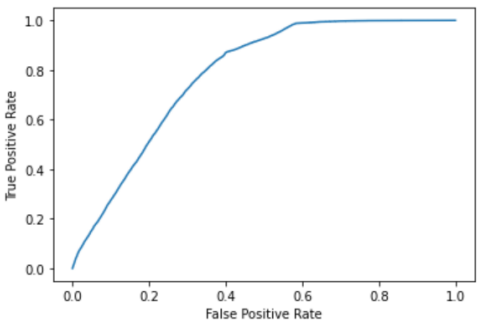
Blow are the performance measures:

```
Accuracy Using Logistic Regression :  0.88
classification report Using Logistic Regression :
              precision    recall  f1-score   support

           0       1.00      0.88      0.94     91467
           1       0.00      0.00      0.00         0

    accuracy                           0.88     91467
   macro avg       0.50      0.44      0.47     91467
weighted avg       1.00      0.88      0.94     91467

Recall Using Logistic Regression :  0.88
Precision Using Logistic Regression :  1.0
F1-Score Using Logistic Regression :  0.94
Confusion Matrix Using Logistic Regression :  [[80311 11156]
 [    0     0]]
```



ROC Curve for Logistic Regression

KNeighborsClassifier

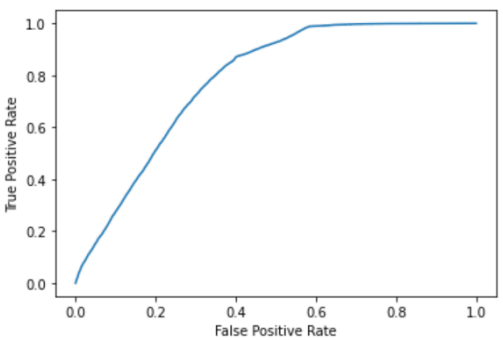
Below are the performance measures:

```
Accuracy Using KNeighborsClassifier :  0.86
classification report Using KNeighborsClassifier :
              precision    recall  f1-score   support

           0       0.96      0.89      0.92     86752
           1       0.13      0.30      0.18      4715

    accuracy                           0.86     91467
   macro avg       0.54      0.59      0.55     91467
weighted avg       0.92      0.86      0.88     91467

Recall Using KNeighborsClassifier :  0.86
Precision Using KNeighborsClassifier :  0.92
F1-Score Using KNeighborsClassifier :  0.88
Confusion Matrix Using KNeighborsClassifier :  [[76992  9760]
 [ 3319  1396]]
```



ROC Curve for KNeighbors

RandomForest

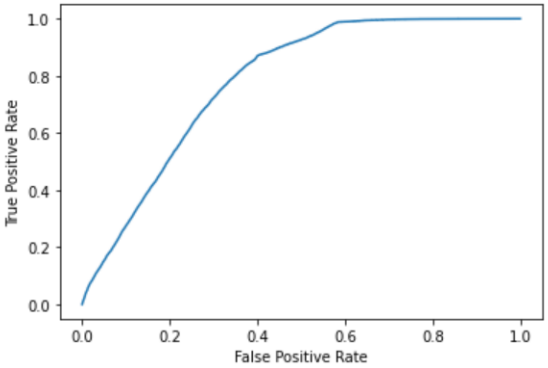
Below are the performance measures:

```
Accuracy Using RandomForestClassifier :  0.88
classification report Using RandomForestClassifier :
              precision    recall  f1-score   support

           0       1.00      0.88      0.94     91467
           1       0.00      0.00      0.00         0

    accuracy                           0.88     91467
   macro avg       0.50      0.44      0.47     91467
weighted avg       1.00      0.88      0.94     91467

Recall Using RandomForestClassifier :  0.88
Precision Using RandomForestClassifier :  1.0
F1-Score Using RandomForestClassifier :  0.94
Confusion Matrix Using RandomForestClassifier :  [[80311 11156]
 [    0     0]]
```



ROC Curve for RandomForest

RESULT AND DISCUSSION

According to the performance evaluated and mentioned in the previous topic, it can be noted that Logistic Regression and RandomForest Classifier gives the highest accuracy i.ie.;88% of the trained model. This is because RandomForest Classifier algorithm is less prone to overfitting and has higher efficiency as compared to other algorithms.

Additionally, the Reciever Operating Charateristic Curve (ROC) of RandomForest Classifier algorithm is close to 1. This concludes that RandomForest is the most efficient algorithm out of the three than have been chosen.