**Abstract**

**The Employees of Private and Governments sector to have Travel Insurance are most important necessary. The main objective of this research work is to identify meaningful and decisive factors for travel insurance in a learning context through exploratory data analysis (EDA) and feature selection techniques. Also, machine learning algorithms are applied to the datasets and are evaluated using performance metrics.**

**Keywords:** (*Travel Insurance data, Classification, Exploratory data analysis Feature selection, Machine learning algorithms).*

**1.Introduction**

Insurance is financial protection against any type of risk. This protection is defined by a contract between two parties: the insurer and the insured (beneficiary).In case of an unforeseen event, the insurer must pay a claim to the insured i.e. the benefit amount to be paid to the beneficiary according to the policy document. Depending on the types of risk involved, insurance policies are of different types. Life Insurance, Travel Insurance, Health Insurance, Auto Insurance and Property Insurance are some of the different Line of Business’ (LOB) in the insurance industry.

Another type of insurance is Re-insurance in which an insurance companies & government purchases insurance from another insurance company to protect itself from financial risk due to the event of an enormous claim.

The whole industry runs on the concept of risk or financial loss reduction. In such a deal, where the insurer has to ensure the client against any type of financial loss against any unforeseen event, there must be a way for the insurer or the insurance company to manage their transactions to pay claims as well as generate sufficient profit to survive in the industry. Now days this term used in association with bringing new technologies and innovations in the field of insurance to impact the regulatory practices of insurance markets. Here comes the concept of Machine learning. Due to the advancement of technologies, the volume of data to be dealt with has enormously increased and this data can be structured, semi-structured as well as unstructured. Data Science technologies help to analyze data and extract important information from it that can lead to better decision making and strategic business development as compared to the above-mentioned traditional data-processing techniques. Along with Data Mining, the industry is shifting towards machine learning algorithms to predict using the analysis done on big datasets for better detection, The use of machine learning algorithms in this process to helps which employees have already take insurance or not.

The purpose of this research is to understand machine learning algorithms helps which employees have already can take insurance or not various sector such as government/private.

In this analysis, datasets are used to perform predict travel insurance of employees analysis using different classification algorithms. Four feature selection algorithms have been used to predict the dimensionality of the data and to improve the results of the analysis. The algorithms are finally evaluated and compared based on four widely approved and trusted metrics: accuracy, precision, recall and f1-score.

**2. Data Visualization**

Data visualization is the phase of machine learning in which data is explored and studied in depth. The hidden features are analyzed. To begin in this phase Exploratory Data Analysis has been done on the data. It helps to draw meaningful insights from the data in the form of statistical and graphical representation.

**2.1 Analysis of target label**

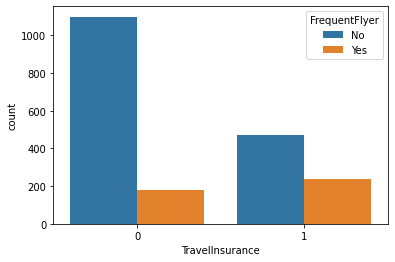
Beginning with target value distribution over attributes present in the data.

**Figure.1 Target Labels**

Infringes that can be drawn from fig.1 tell, “0” class represents: the customer will not leave the company service. “1” class represents: the customer will leave the company service. More instances i.e almost 85% are of Churn category that means customer will leave the company on the basis of the attributes mentioned.

**2.2 Analysis of attributes**

Count of attributes distributed over the dataset.



**Figure.2 Count plot for Travel Insurance**

The employment data frequent flyer No attributes is more as compare to Yes attributes.

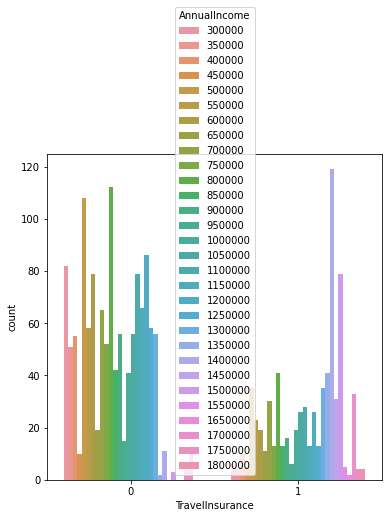
On the basis of data Travel insurance provided by Government and private sector,

The frequent flyer employees

**2.3 outlier Detection**

Box plot helps to determine if there are outliers present in the attributes in a given dataset

**2.4 Paired analysis**



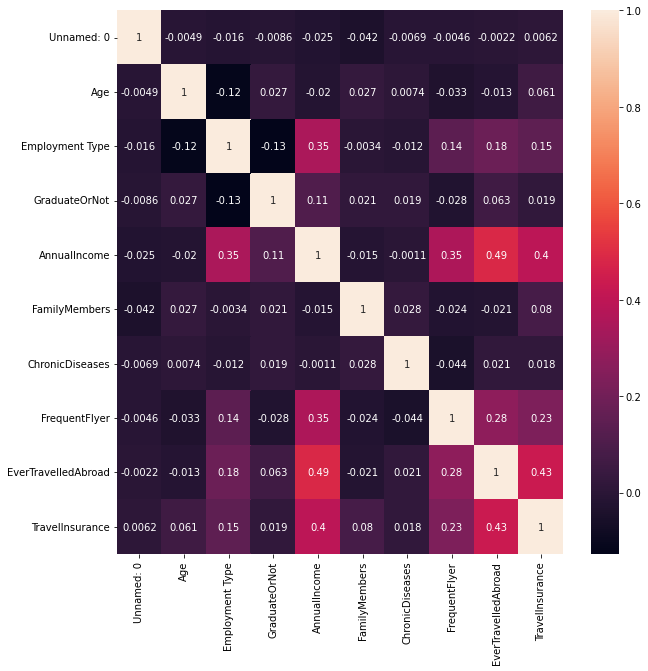
**Figure.4 Plotting Mean Value for Churn**

**2.5 Statistic Visualization**

Analyzing mean corresponding to each attribute with respect to travel insurance.

**2.6 Attributes correction**

Determining correlation between independent and dependent features contained in the dataset.



**Figure.5 Correlation Matrix**

From fig.5 some features can be seen highly correlated with each other where the correlation present is greater than 50%. The dimensionality reduction can be performed to overcome this.

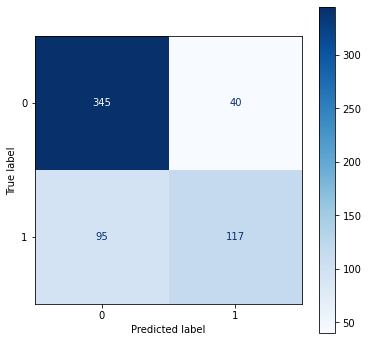
**3 Experiments**

**3.1 Predictive Machine Learning Models**

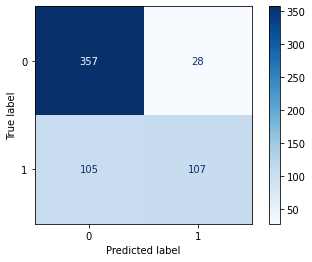
In this project, four machine learning models are used for classification. The dataset with labeled classes are given with the inputs during model learning.

**3.1.1 K Neighbors Classifier**

The k neighbor algorithm is greedy in nature. It is the basic classification algorithm with a quick implementation but it misses the data points which are at shorter distances. It works on a similarity method. Various distance functions are used for finding the values of k. In this work. Optimal value of k has been obtained by observing the score at each k, taking values of k from 1 to 29. The best K value has come out to be 5 with maximum accuracy score achieved as 92.0 at this particular K. Increase in value of K, further, leads to decrease in accuracy score reflecting not to choose higher K value to prevent a huge degradation in model performance.

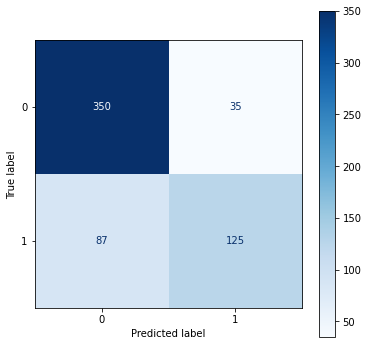


KNN with hyper parameter tuning

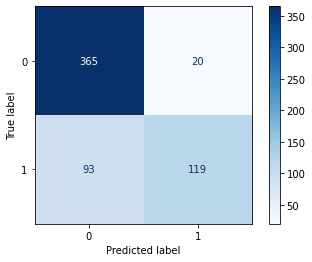


**3.1.2 Random Forest**

It is supervised predictive learning which follows the divide and conquers method to split the data, which is represented by the nodes. The trees formed are trained on the dataset randomly and. The final output generated is the output based on the output produced by the maximum number of decision trees. Random forest classifier is capable of dealing with a large number of attributes as well as missing data points. The subplot curve obtained for random forest classifiers is shown in the figure below.



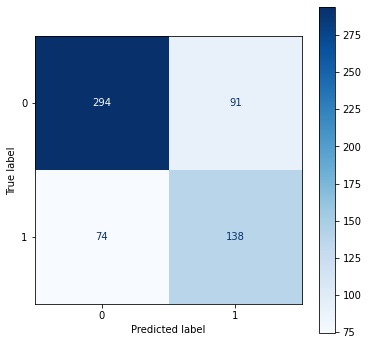
**RF with hyper parameter tuning**



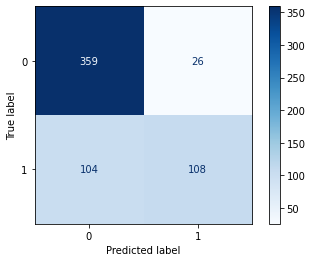
**3.1.3 Decision Tree**

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.



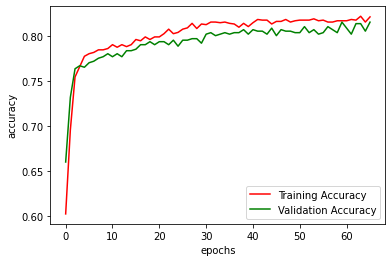
**DT with hyper parameter tuning**



**3.1.4 Artificial neural network**

The model with the highest Accuracy Score is Random Forest Classifier. At the same time, when looking at the subplot curve of Random Forest Classifier, it is seen that it learns the classes better than other models.

layers have been used in this model. The output layers produce the classes of new observations.



The plot shows the total accuracy and validation accuracy at each epoch compiled during training and validation of the neural network. The accuracy during training and validation chasing each other with a similar pace, resulting in average accuracy as 80 and 82 on training and validation respectively.

**3.2 Feature Scaling**

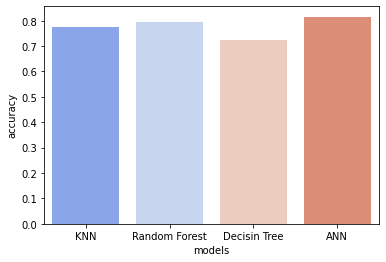
Feature scaling is the data processing step, used to standardize the existing independent attributes in the dataset in a definite range. The need of scaling the data points comes into picture when values are deviated. Standard scaler is applied here as a scaling technique. The values are then surrounded about the mean with a fixed standard deviation.

**3.3 Dimensionality reduction**

**4 Results**

Comparison of performance

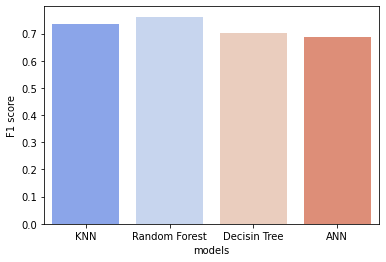
To evaluate the model performance, different performance metrics have been used, such as AUC subplot score, Accuracy score, Cross validation score, Confusion Matrix on test data. With the help of a confusion matrix, the number of correctly classified classes can be calculated. The high AUC score indicates better model performance on positive as well as negative classes. The cross validation score achieved by KNN and Random Forest are 0.7338 and 0.7956 respectively, Decision tree and ANN are 0.7236 and 0.8157



**Figure.8 Performance Plot of All Models**

**Table.1 Performance Metrics of All Models**

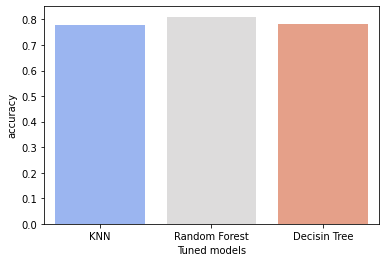
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-Score |
| K Neighbor | 0.7738 |  |  | 0.7352 |
| Random  Forest | 0.7956 |  |  | 0.7618 |
| Decision Tree | 0.7236 |  |  | 0.7033 |
| ANN | 0.8157 |  |  | 0.6857 |



F1 Scores:

--------------------------

Performance comparison of tuned models



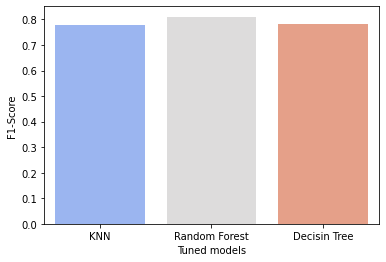
Tuned Accuracy Scores:

--------------------------

KNN : 0.7738693467336684

Random Forest : 0.7956448911222781

Decisin Tree : 0.7236180904522613



Tuned F1-Scores:

--------------------------

KNN : 0.7352549889135254

Random Forest : 0.7618122596342516

Decisin Tree : 0.7033634170799794

**5. Discussion**

**5.1 Dataset**

The dataset taken for Employment Travel Insurance prediction is from the telecom industry picked from kaggle repository. The link to dataset is provided as:

The dataset is composed of approximately 1987 instances and 10 features. It is free from missing or null values. The attributes contained in the data are of numeric type as in integer or float. To scale down the data points Standard scaler is applied.

5.2 conclusion

5.3 Future work

6.Referencces