University of Essex
School of Computer Science and Electronic Engineering
CE802 Machine Learning and Data Mining

CE-802
Machine Learning and Data Mining

Assignment: Design and Application of a Machine Learning System for a Practical Problem**.**

Report on the Investigation

Created By:

Name: Srikanth Reddy G

Reg.No: 2004445

Word Count: 1100+

This report comprises of two main sections, one, where I will discuss about the techniques used for building the Classification algorithms and its results, and the second, where I will explain about the Regression task.

The objective of the first task is to predict weather a customer would claim the insurance so that discounted insurance quotes could be given for the ones that have very less chance of claiming. Since we already have their past records which tells about their claim history, we will be using Supervised Learning technique to identify the two distinct customer segments.

Classification Task:

- **Pre-Processing Phase**

The data comprises of 15 distinct features and a dependent class. Out of all the available features, the last feature, 'F15', contains almost half missing values. When we have a feature where almost half of the values are missing, it is good to discard it. It becomes very difficult to impute the values since the feature's distribution change with a big margin. In the experiment, I have done the median imputation and I have trained the algorithms with and without the F15 and compared the results. It looked like the imputation method seemed to be working as it had given high accuracy. The

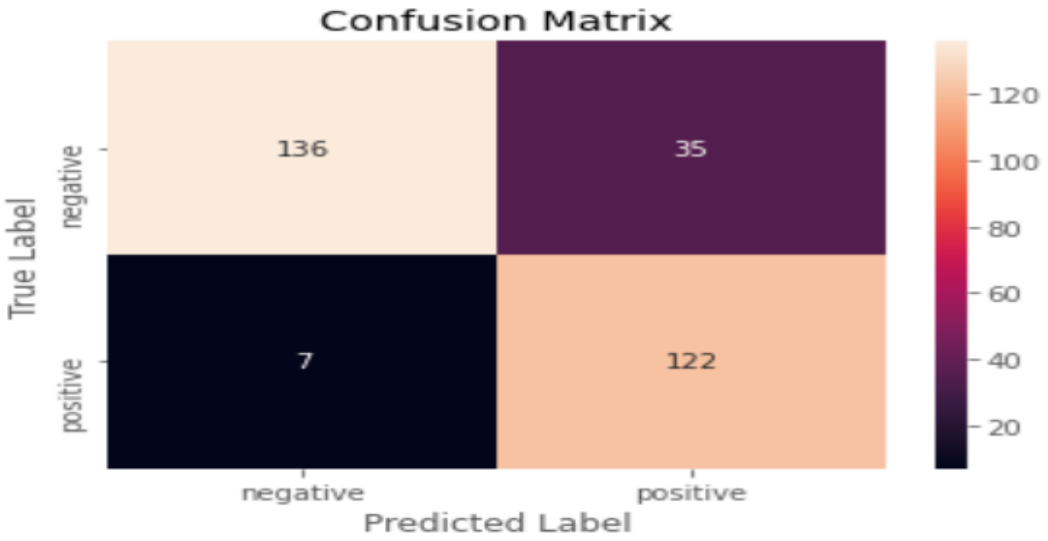
- **Model Building**

For training the model, 80% of the data has been taken and the rest for testing. The first model that the data has been trained with is the Decision Tree algorithm. Decision Tree is known for yielding high accuracy models given the dimension of the data is less, which is true in our case. Feature Scaling is an important step which should not be missed when dealing with numeric data, but Decision Tree does not require the scaling for training. The model has been tuned with different parameters using Random Search CV method before testing on the testing set. The second algorithm which was used for training is Logistic Regression. It is one of the best algorithms which gives you highly accurate probabilistic results if

tuned correctly. With the correct 'C' and Rgularization parameters, it produced second best model in terms of accuracy and Recall. The third and final model used in this case study is SVM. Insurance is a domain where we are expected to have a lot of features since we want to know a lot about the customers. This could create 'Curse of Dimensionality' and a lot of models do not perform well in such case. However, SVM triumph over such scenarios where it transforms the given features into higher dimensions and finds the correct hyperplane. The trained SVM model has been proven to make the best predictions for the given data.

- **Model Evaluation**

Now that the data is trained, we must test it making live predictions on the new dataset. Accuracy may not be the best way to prove the performance. Metrics like Confusion Matrix, Recall Score, Precision have been used for the evaluation process.



SVM's Confusion Matrix

- **Results Interpretation**

All the models were built twice, once with the feature 'F15' and once without the feature 'F15'.

Out of the three Supervised Learning algorithms, SVM produced the best results followed by Logistic Regression and Decision Tree in both the instances. We are using all the metrics to evaluate the performance with more biased towards the Recall Score. The reason for concentrating more

on the Recall Score is because we want a model to have very less False Positive Rate (FPR). In the field of Insurance, we do not want a bad customer to be identified as a good customer.

	Algorithm	Accuracy	Precision	Recall	F1-Score
0	Decision Tree	75.66%	0.73	0.69	0.709
1	Logistic Regression	77.333%	0.667	0.946	0.782
2	Support Vector Machine RBF	81%	0.71	0.93	0.805

Results of Models built without F15 feature.

	Algorithm	Accuracy	Precision	Recall	F1-Score
0	Decision Tree	80.667%	0.775	0.775	0.775
1	Logistic Regression	83.667%	0.753	0.922	0.829
2	Support Vector Machine RBF	86%	0.777	0.946	0.853

Results of Models built with F15 feature

We can clearly see that Missing value imputation has proven to be working. Now we need to see, which algorithm to choose for making prediction on the unseen dataset. We can select SVM since it has the highest Accuracy, Precision and Recall score.

Regression Task:

- **Problem Statement:**

Now that we have classified the customers into 2 segments, we want to identity the customer's claim amount. We will apply regression algorithms to find out the claim.

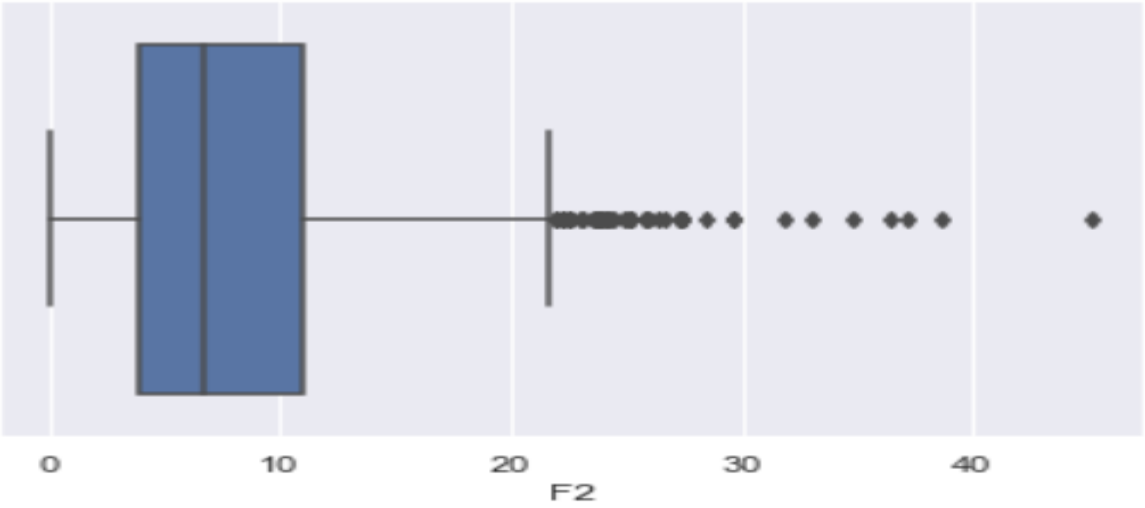
- **Pre-Processing Phase:**

  We have observed categorical variables in the data and unfortunately, Regression algorithms cannot deal with categorical variables. We must encode the variables into numeric features. We will use the 'get_dummy' function to convert the values into new features.

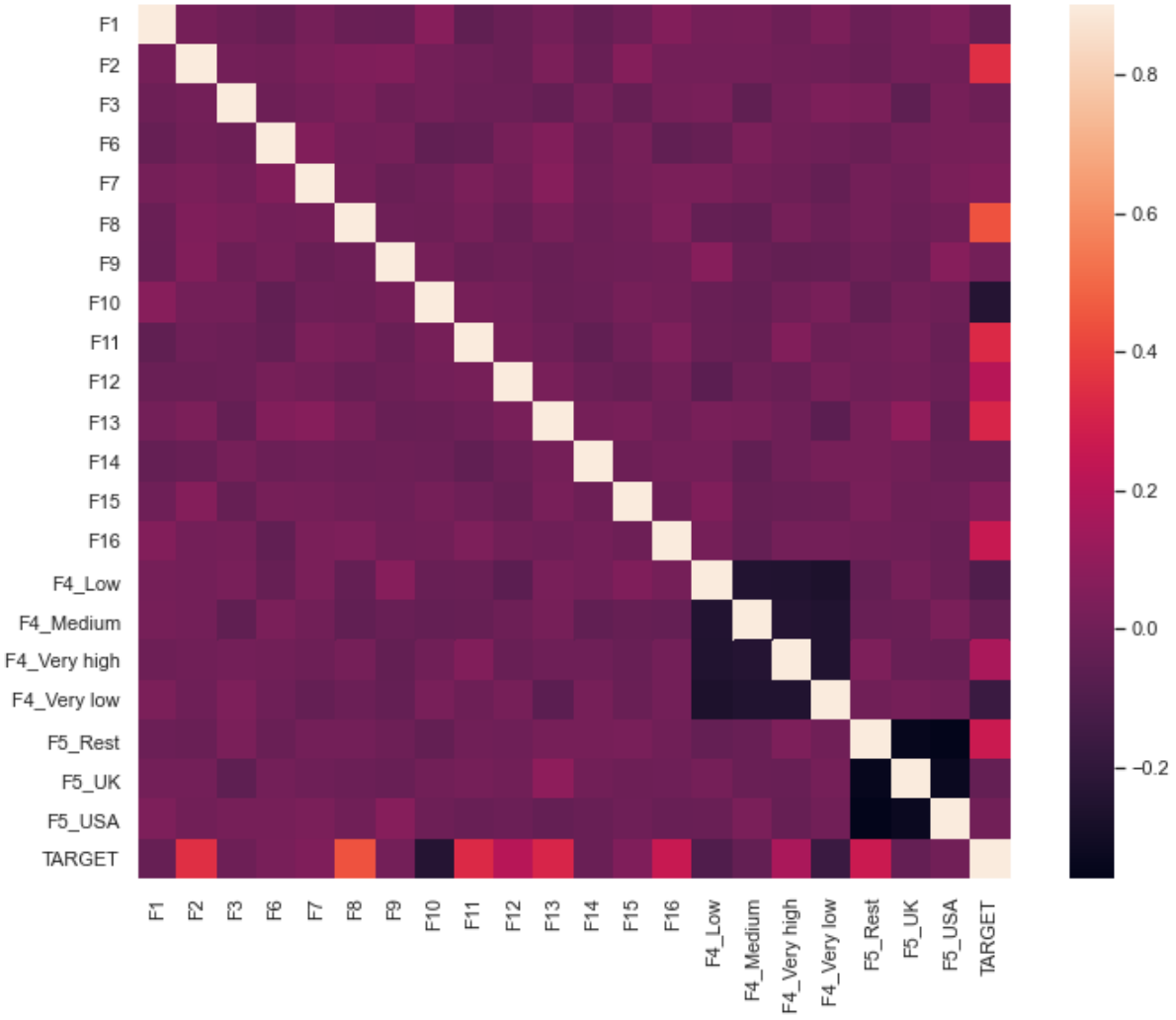
F4_Low	F4_Medium	F4_Very high	F4_Very low	F5_Rest	F5_UK	F5_USA	TARGET
0	0	0	0	0	1	0	1051.99
0	1	0	0	0	1	0	816.64
1	0	0	0	1	0	0	3241.77
0	0	0	0	1	0	0	0.00
0	1	0	0	0	0	1	0.00

In the above diagram, we can see that the values of the categorical variable have been encoded as new features.

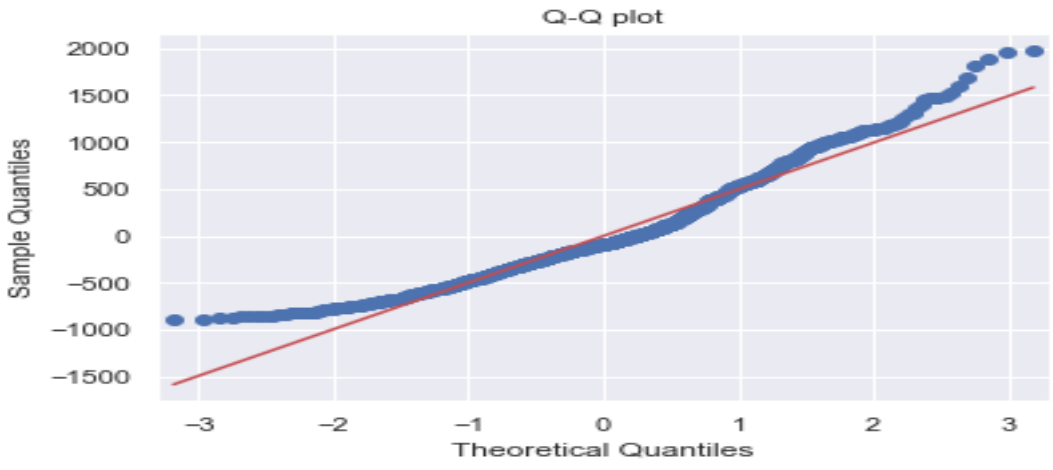
Regression problems only work if all the regression assumptions are true. Few tests have been conducted to make the data perfect of training. In the first test, the outliers have been eliminated as they may pull the correct Regression line away from the desired points. Box Plots have been plotted to check if the data has outliers and then with the help of Z-Test, outliers have been removed.



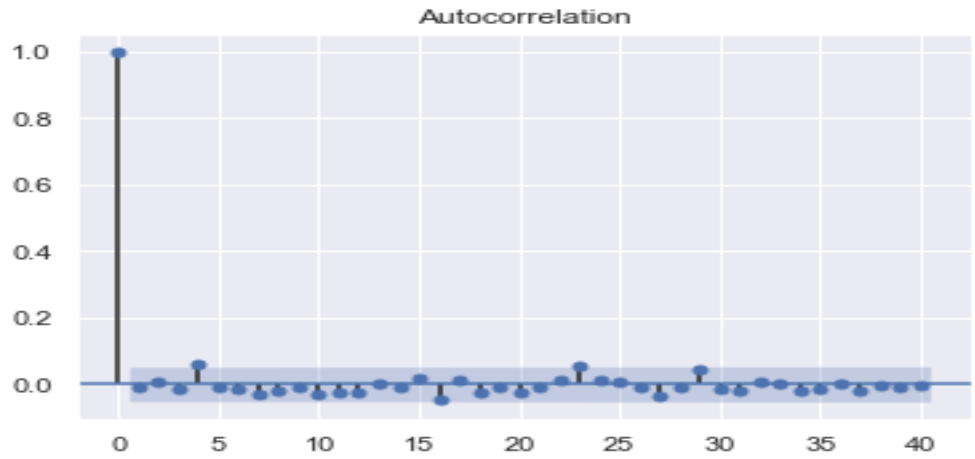
Now that the data is free from outliers, we will check for the collinear features. We must eliminate the features that are collinear with each other. A correlation plot has been plotted and VIF is also calculated to check if Multi-Collinearity issue occurs.



A Q-Q Plot of Residuals are also plotted to check if they are distributed Normally. In this case, we can see that all the Residuals are not on the line. It means, the Residuals are not distributed Normally. One of our assumptions failed. One of the solutions for such problems is to apply Feature Engineering hacks and transform the data.



Auto Correlation Check has been performed since we don't want any kind of relation between the Features and the Residuals. In this case, we can see that there is no presence of Autocorrelation.



- **Model Training**

All the required statistical test has been done before the model training. Three algorithms have been selected for training the data, Linear Regression, Decision Tree Regressor and Random Forest Regressor. Out of the three algorithms, Linear Regression has proven to be the best one.

- **Tuning the Models**

After the data has been trained by the Linear Regression Algorithm, Backward Elimination technique has been applied to remove the features with a P-Value more than 5%. This method didn't work since the models didn't show any improvement in the R^2. In the similar manner, Random Forest, and Decision Tree Regressor have been trained and tuned with different parameters.

- **Result Analysis**

It can be clearly seen that the MSE of Linear Regression is the least of all the algorithms. Hence the reason we choose Linear Regression to train the unseen dataset. The remaining algorithms also performed well but it is a proven fact that Tree based algorithms are bad learners if the dataset has too many features. In the Regression task, Categorial Features

transformation has increased the dimensionality of the data, as a result, the Tree Learners didn't work properly.

	Algorithm	R^2	MSE
0	Linear Regression	0.796	295729
1	Decision Tree Regressor	0.6931	655834
2	Random Forest Regressor	0.75195	412177