REPORT

(Word Count: 915)
(Reg Id: 2003258)

Pre-Processing

Part 2

Feature F15 in dataset has a missing data, employed multiple methods to address the issue it, Like, replacing the missing value with mean, regression imputation and entirely discarding the feature. In regression imputation, Linear regression model is trained with features F1 to F14 as input and F15 as target variable. Using the trained model missing values of F15 are predicted.

Part 3

In P3 dataset, there are couple of categorical features, that are F4 and F15. F4 column has 4 unique values (UK, US, EUROPE, REST), as there is not intrinsic order among the values, one-hot encoding is done. Where as in F15 feature, the unique values (Very low, low, medium, high, very high) have an order, so, ordinal encoding is used.

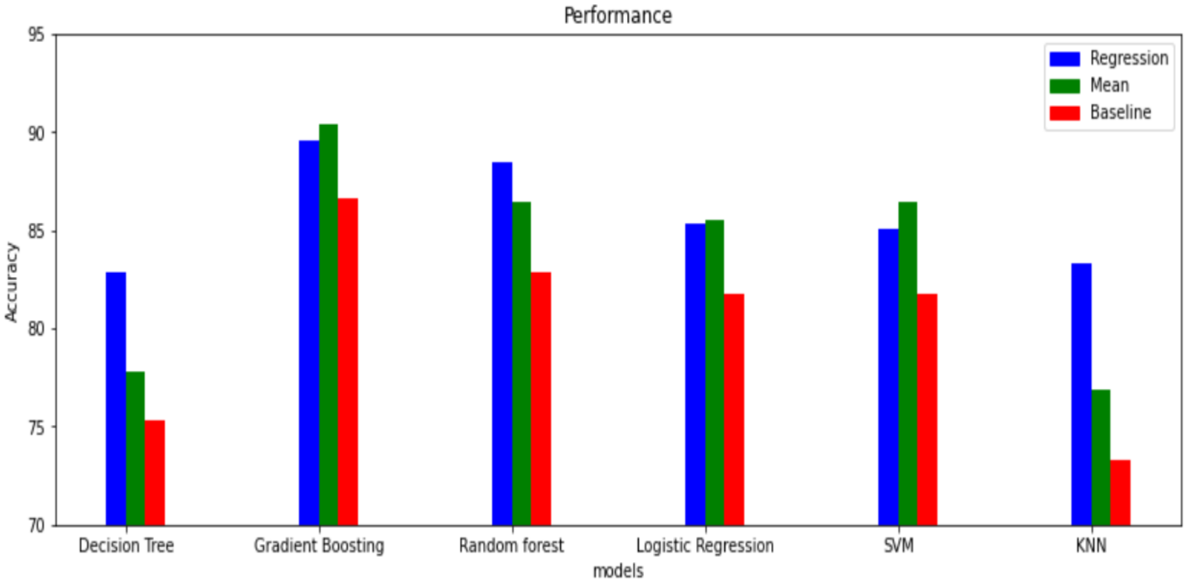
Model analysis:

part2

The models that are used for this task are:
i) Decision Tree,
ii) Logistic regression,
iii) SVM,
iV) KNN,
V) Random forest,
Vi) Gradient boosting.

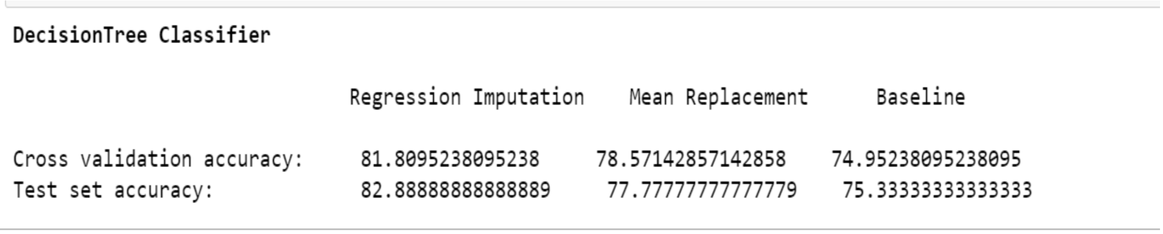
All models are trained and analysed with three different data sets with different imputation techniques.
i)Baseline (Remove Feature F15)
ii) Mean replacement
iii)Regression model (Linear regression)

The performance of various models for a given data set with different imputation techniques are shown below:



Some of the observations we can make from the above plot are, Mostly, performance of all models with baseline imputation technique of removing the whole feature, is significantly less compared to other techniques. Whereas mean replacement technique fared better results in gradient boosting and support vector machines (SVM) while regression model imputation is better in other methods under test. As can be seen from the above plot, Gradient boosting performed better than other models.

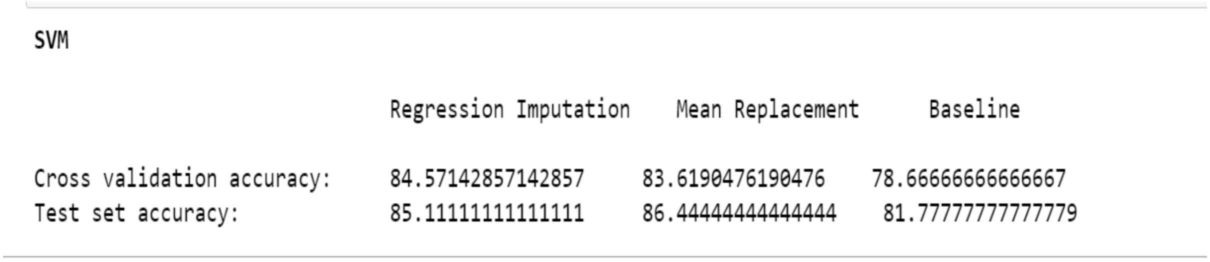
As it is seen the performance plot, Decision tree performance well with regression model imputation technique. It accuracy on test set is 82%. Overall, it is least accurate model compared to others for a given set. It has taken more time compared to methods like logistic regression and KNN. And also has a burden of hyper tuning quite a bit.
Given there are only 15 features, Dt would be much easier to interpret. One benefit of DT is its ability to work with non-scaled data. Logistic regression in this case was needed around 10000 iterations to converge on non-scaled data.



```
DecisionTree Classifier

                             Regression Imputation    Mean Replacement      Baseline

Cross validation accuracy:     81.8095238095238     78.57142857142858    74.95238095238095
Test set accuracy:             82.88888888888889     77.77777777777779    75.33333333333333
```

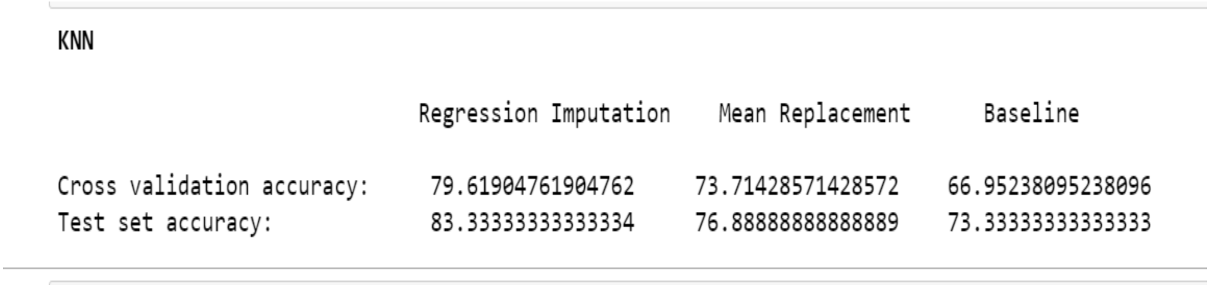
Unlike many other models, SVM performed better with mean replacement imputation. Its accuracy is at 86%, while techniques like regression and baseline approach are at 85% and 82% respectively. For a given data set, SVM with linear kernel is performing better than nonlinear kernels, almost 10% increase in performance is observed.

SVM seems to have done better compared to KNN, logistic regression and Decision tree but overshadowed by Gradient boosting.



```
SVM

                              Regression Imputation    Mean Replacement      Baseline

Cross validation accuracy:    84.57142857142857     83.6190476190476    78.66666666666667
Test set accuracy:            85.11111111111111     86.44444444444444    81.77777777777779
```

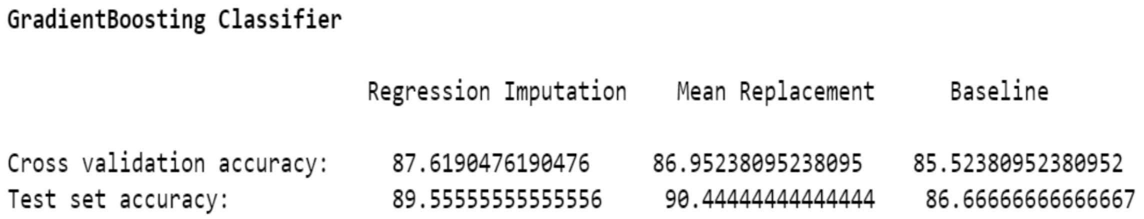
KNN seemingly, one of the worst performers among the tested models. Its accuracy is at 83% on test data on using regression imputation, there is an increase in the performance on using regression imputation compared to other models.



```
KNN

                              Regression Imputation    Mean Replacement      Baseline

Cross validation accuracy:     79.61904761904762     73.71428571428572    66.95238095238096
Test set accuracy:             83.33333333333334     76.88888888888889    73.33333333333333
```

Gradient boost performed slightly better on mean imputation approach compared to Regression and baseline approach down by few points compared to others. Its accuracy on test data set is highest among all models at 91%. Training model is taking significantly more time consuming which is also the case with hyper tuning.

Looks like additive modelling helps gradient boost to approximate almost but not entirely linear data set where algorithms like logistic regression and SVM fall short off.



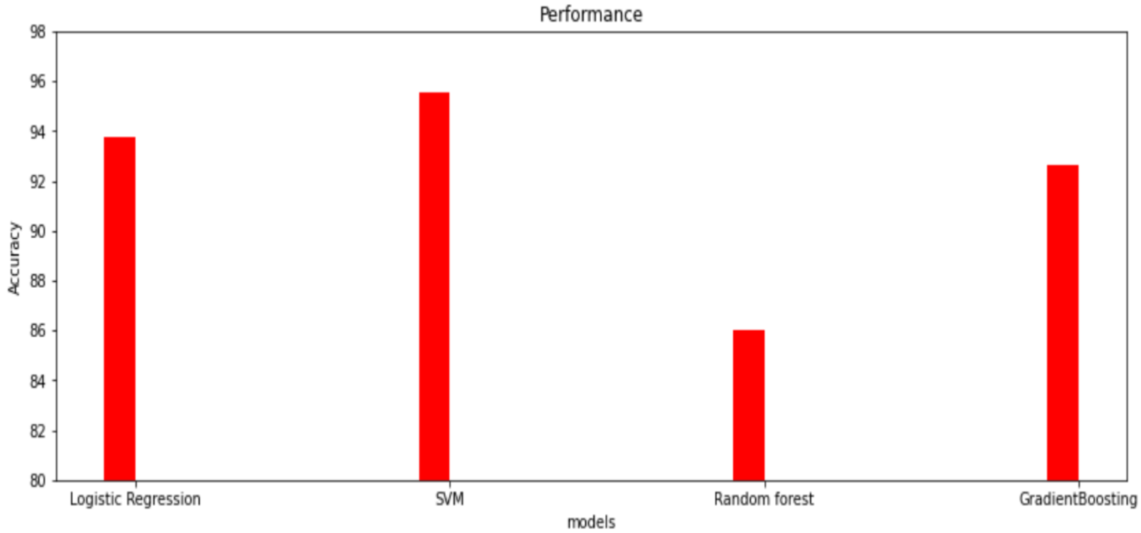
```
GradientBoosting Classifier

                          Regression Imputation    Mean Replacement      Baseline

Cross validation accuracy:    87.6190476190476     86.95238095238095    85.52380952380952
Test set accuracy:            89.55555555555556     90.44444444444444    86.66666666666667
```

Part 3

Here, we need to first predict, if insured would make a claim, and then if he does, how much?
We can divide problem into two parts classification part and regression part.
In order to make training data for both the problems, we divide data in two ways.
One dataset with Boolean target variable for classification problem and other with target variable with amount claimed by insured for regression part.
Once Both the models are trained, we can first use a classification model to predict if insured would make a claim in future or not. If insured would not probably make a claim, the Target value would be 0. Otherwise, trained regression model is used to predict the target variable.

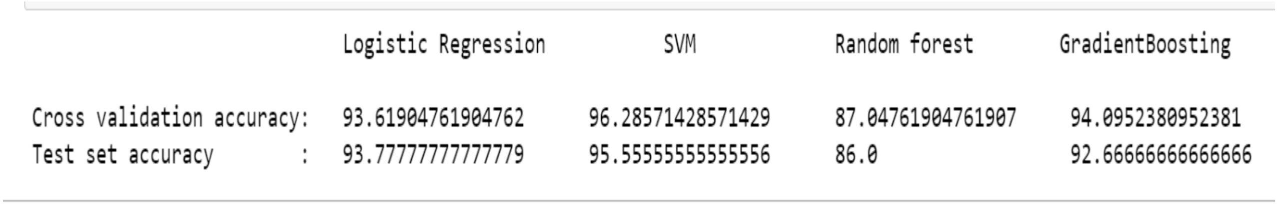
Here, we use both classifier modes and regression models to predict the target value.

The performance of classifiers used is as shown as below



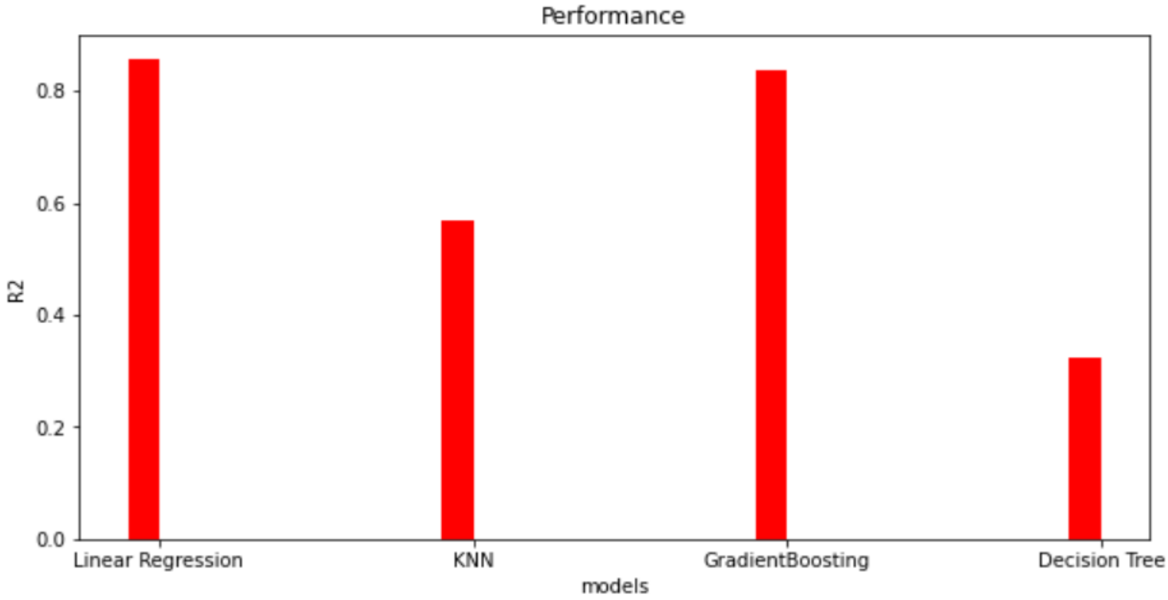
For the given data set, as can be seen in the plot above, SVM with linear kernel is producing the highest accuracy at 95%, Followed closely by Logistic regression and gradient boosting at 94% and 93% respectively. The performance of linear kernel is significantly higher than the nonlinear kernel and also logistic regression is performing well, that means, data is linearly separated.
The performance of random forest is worst among all 4 models at 86%.

Below are the accuracies obtained by different classifiers for a give data:



```
                             Logistic Regression        SVM            Random forest        GradientBoosting

Cross validation accuracy:   93.61904761904762    96.28571428571429    87.04761904761907    94.0952380952381
Test set accuracy        :   93.77777777777779    95.55555555555556    86.0                 92.66666666666666
```

Below are the r2 values obtained for various regressors for insurance claim prediction

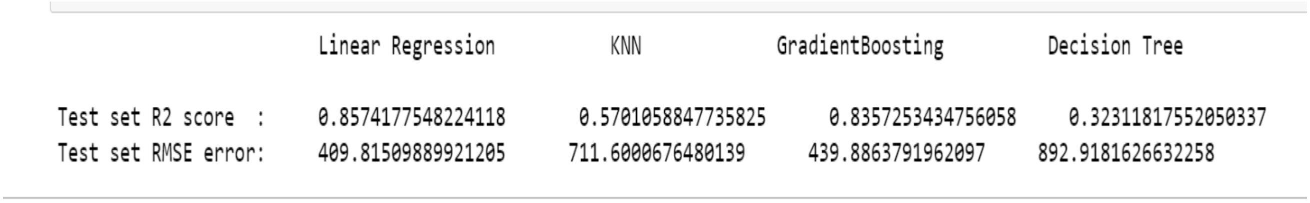


Second part of P3 includes predicting the possible claim if insured is going to claim the insurance. Various models like linear regression, KNN, Gradient boosting and Decision Tree are trained and compared.

Metrics like R2 and RMSE error are used for model accuracy.

As can be seen in the above plot, Linear regression outperformed other models like KNN, Gradient boosting and decision tree at R2 0.86 and RMSE 409, and closely followed by Gradient boosting at 0.84 and 439 (R2 and RMSE respectively).

Like, seen in different problem statements before, Decision tree largely under performed with 0.32 and RMSE 893.While KNN performance is consistent across various K values at 0.57 and 711 (R2 and RMSE respectively)



```
                       Linear Regression        KNN              GradientBoosting         Decision Tree

Test set R2 score  :   0.8574177548224118    0.5701058847735825    0.8357253434756058    0.32311817552050337
Test set RMSE error:   409.81509889921205   711.6000676480139    439.8863791962097    892.9181626632258
```