Member:

Tianzhi Cao (tc324@njit.edu)

File Descriptions:

Admission_Predict_Ver1.1.xlsx

Introduction of This data:

This dataset is inspired by the UCLA Graduate Dataset. It include 500 students GRE and TOFEL Score, university ration, SOP strength, GPA, research experience and their chance of admit. The test scores and GPA are in the older format. The dataset is owned by Mohan S Acharya.

Data Fields:

0. Student Serial Number

1. GRE Scores (out of 340)

2. TOEFL Scores (out of 120)

3. University Rating (out of 5)

4. Statement of Purpose Strength (SOP) (out of 5)

5. Letter of Recommendation Strength (LOR) (out of 5)

6. Undergraduate GPA (CGPA) (out of 10)

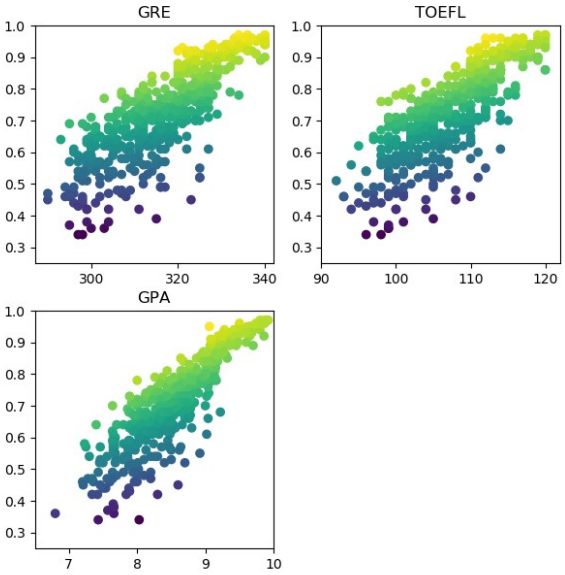
7. Research Experience (either 0 or 1)

8. Chance of Admit (ranging from 0 to 1)

Tasks:

1. Visualize data
    - Visualize the relationship between GRE grade and Chance of Admit
    - Visualize the relationship between TOFEL grade and Chance of Admit
    - Visualize the relationship between GPA and Chance of Admit

    Run drewgraph.py to get the graphs, as we see, the points on GPA graph more concentrated. Which mean GPA has more relationship with chance to be admit.



2. Outliers
    - Using boxplot find outlier points for GRE/TOFER/GPA and Chance of Admit

    Outliers were defined: bigger than (IQR3-IQR1)+1.5 IQR or less than (IQR3-IQR1)-1.5 IQR. There are two outlier points for chance to be admit but no outlier point for GPA/GRE/TOEFL. Run outlier.py get this result:

```
W:\hw\CS301\Finalproject>python outlier.py
outlier check for Admission Percentage
IQR1 is 0.63
IQR3 is 0.82
Outlier point: 93, value is 0.34
Outlier point: 377, value is 0.34
outlier check for GRE score
IQR1 is 308.0
IQR3 is 325.0
outlier check for TOEFL score
IQR1 is 103.0
IQR3 is 112.0
outlier check for GPA
IQR1 is 8.1275
IQR3 is 9.04
```

3. Modeling
    - Build 2 different models that uses all data from the Admission_Predict_Ver1.1.xlsx

    I used linear regression and decision tree regression.

4. Validation
    - Perform 70-30 holdout and present MSE, MAE, and R-Squared for both of two model

```
MSE test for linear_model: 0.003176376379901243
MSE test for DecisionTreeRegressor: 0.007201333333333334
MAE test for linear_model: 0.026764621537002597
MAE test for DecisionTreeRegressor: 0.040000000000000036
R-squared test for linear_model: 0.8273831906033338
R-squared test for DecisionTreeRegressor: 0.6086511689019336
```

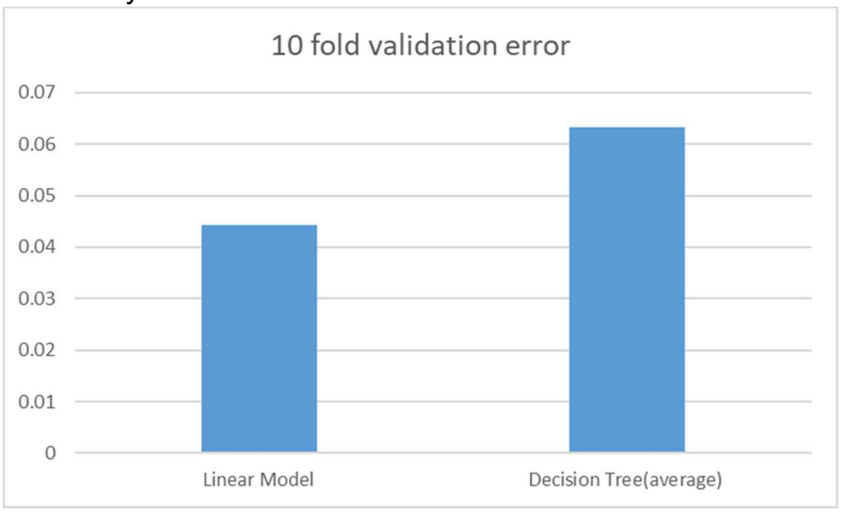
    - If we say 75% is the line to decide if a student should apply this university, get precision, Accuracy and Recall for these two models.

```
If we say 75% is the line to decide if a student should apply this university.
Precision for linear model: 0.9041095890410958
Precision for Decision Tree Regressor: 0.8051948051948052
Recall for linear model: 0.88
Recall for Decision Tree Regressor: 0.8266666666666667
Accuracy for linear model: 0.8933333333333333
Accuracy for Decision Tree Regressor: 0.8133333333333334
F1 for Linear Model: 0.8918918918918919
F1 for Decision Tree Regressor: 0.8157894736842106
```

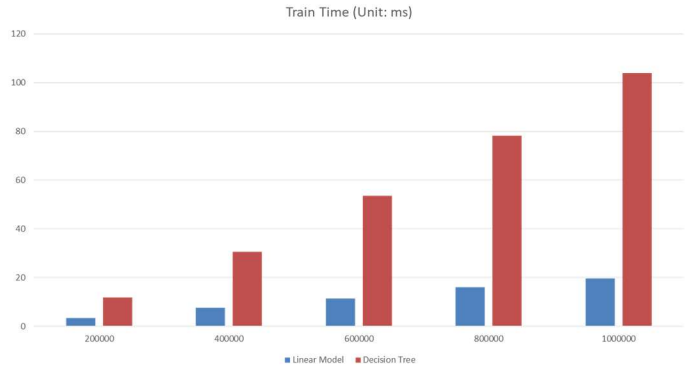
5. Compute confidence interval of Model 1 and Model 2 for the following different confidence levels: 80%, 90%, 95%
   There are no significant different between these two models in confidence levels 80%, 90% and 95%

```
t value is 330.6138961745208
There is no significant different in confidence level 95%, 90%, and 80%
```

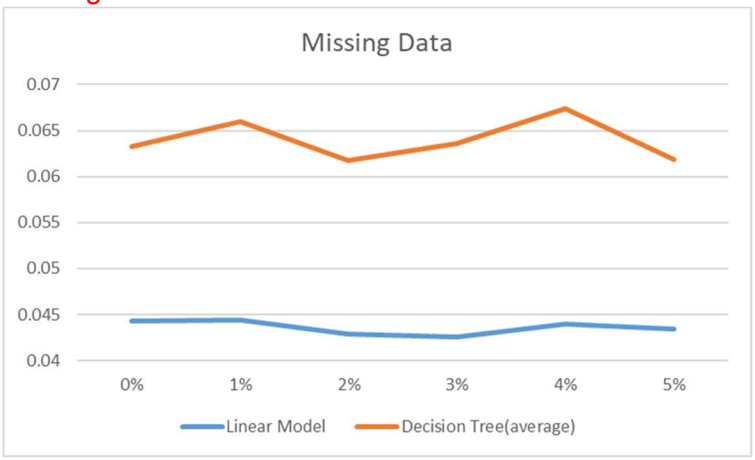
6. Compare these two models by considering
    - A. error by 10-fold validation



- B. efficiency in training time (scalability)
  I increased the size of database to make it has 0.2/0.4/0.6/0.8/1 million data because if there are not enough data in database, time change is implicit.



- C. Robustness
  Missing value test



Add/delete attribute test