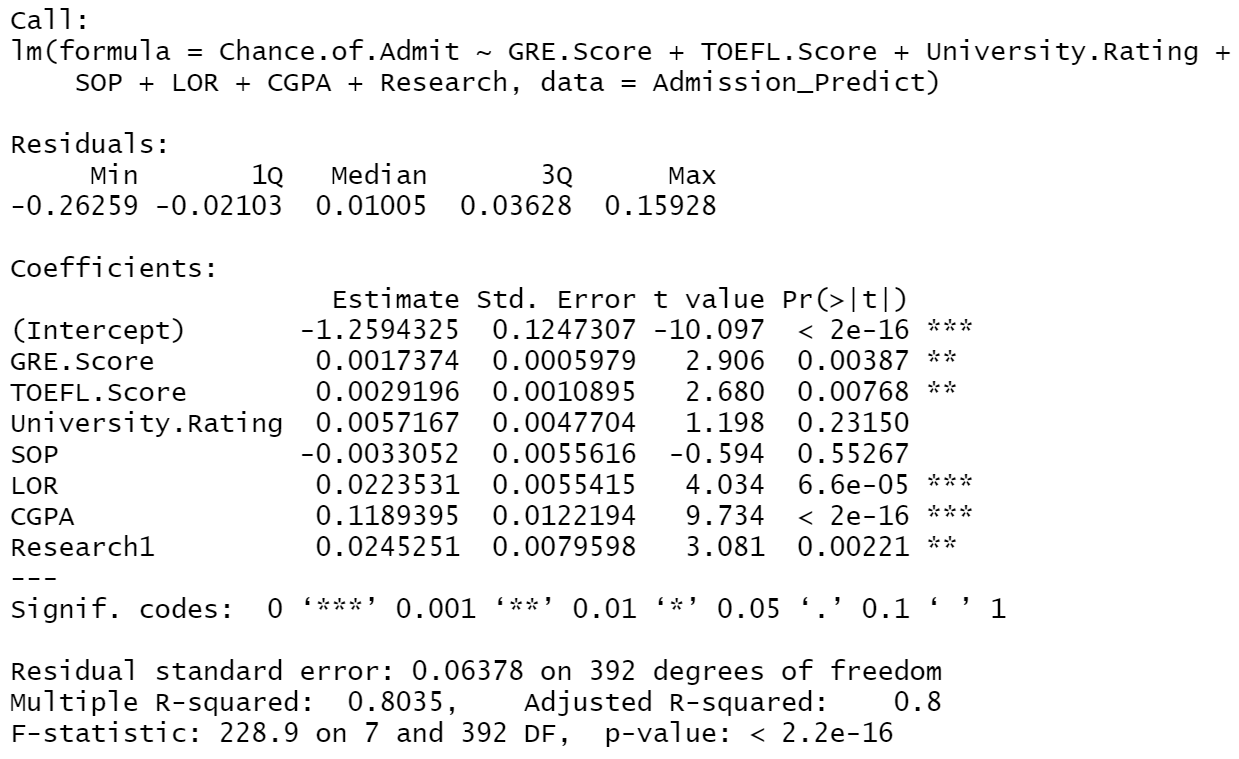
**MATH 4322 Project Final Report - Graduate Admissions for Masters Program**

1. Overall intro, including: **(Paul Duong, Kyle Schulte)**
   1. The data set is called Admission\_Predict and the purpose of this data set is to predict the probability of an individual being admitted to a master’s program given the predictors GRE score, TOEFL score, University rating, Statement of Purpose and Letter of Recommendation Strength, Undergraduate GPA, Research Experience. The data set also has a Serial Number, but this is strictly to identify the university.

The predictors are broken down such as:

* + 1. Serial number: The index for data set containing ID of student
    2. GRE score: A quantitative variable that is out of 340 (General Test Score)
    3. TOEFL Score: A quantitative variable that is out of 120
    4. SOP: A quantitative variable out of 5
    5. LOR: A quantitative variable out of 5
    6. CGPA: A quantitative variable that is out of 10
    7. Research: A quantitative variable
    8. Chance of Admission: The response variable for our data set
  1. The questions/tasks you are posing about this data (main excerpts from your “Formulating Questions/Models” project assignment). (Note: AVOID asking about different response variables)
     1. Can the student’s “Chance of Admit” be effectively predicted with only “CGPA,” and if so what is the significance that “CGPA” has on “Chance of Admit”?
     2. Can the student’s “Chance of Admit” be effectively predicted with “TOEFL Score,” “University Rating,” and “SOP”?
     3. Does the student’s “Chance of Admit” increase if they have “Research” experience, and if so, how much of an impact does it have?

1. Main Overall Question and Results: **(Paul Duong)**
   1. The main question that we will be primarily focused on is:
      1. Can the student’s “Chance of Admit” be effectively predicted with only “CGPA,” and if so what is the significance that “CGPA” has on “Chance of Admit”?
      2. The two models that we will be using to analyze the “Admission\_Predict” data set are the linear regression model, and the decision tree model. We decided on the linear regression model because our response variable “Chance of Admit” is a quantitative value with values between 0 and 1, and the linear regression model is very suitable for a quantitative response variable. It is very easy to implement the linear regression model as well as interpreting the output of the coefficients. When we perform the linear regression model, we are assuming that the response and predictors have a straight-line relationship, but if necessary we can perform transformations such as poly, log, or sqrt to get a better representation on our dataset. We decided to go with the regression decision tree method because not only does it provide us a good visual representation on the data, it also allows us to manipulate the trees into a subset of the tree. By using cross validation, we can get more accurate and improved results. Although it provides a better interpretation of the data, decision tree models struggle with prediction accuracy and suffers from high variance from having subsets of trees that could lead to completely different results.
2. For Linear Model **(Kyle Schulte)**:
3. The model formula is Chance of Admit = β0 + β1(GRE Score) + β2(TOEFL Score) + β3(University Rating) + β4(SOP) + β5(LOR) + β6(CGPA) + β7(Research)
4. Without worrying about training or test data set, I used R studio to fit the model above, and then interpreted it by using the summary function with the following results. 

For most of the values coefficients, the p value was extremely close to zero allowing us to reject the null hypothesis. For the two predictors that had higher P values, University rating and SOP, we can consider removing the higher of the two, SOP. In doing so, the accuracy of the fit is increased by a mathematically insignificant amount when comparing several measurements of each fit (Residual standard error, Multiple R squared, Adjusted R-squared, and P value), allowing me to come to the conclusion that it is not worth it to take SOP out of the equation. At this point, I proceeded to double check the data up to this point, and began splitting the data into training and testing sets.

1. At this point, I performed a 80/20% split on the data putting the majority into the training data set.
2. After training the model on the train split of the data, and testing it on the test split of the data, I calculated 3 different model performance metrics
   1. RMSE: 0.0764295

The Root Mean Squared Error (RMSE) measures the average error performed by the model in predicting the outcome for an observation. It is calculated by taking the square root of the Mean Squared Error (MSE).

* 1. MSE: 0.005841468

The Mean Squared Error (MSE) is the average squared difference between the observed actual outcome values and the values predicted by the model

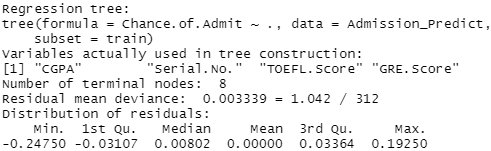
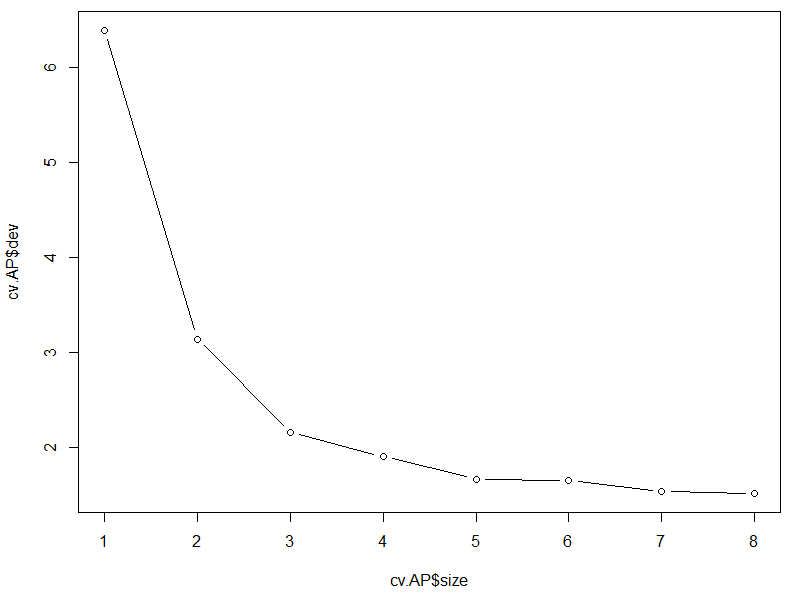
* 1. MAE: 0.04521251

The Mean Absolute Error (MAE) is the average absolute difference between observed and predicted outcomes.

* 1. R2: 0.8025287

The R squared value (R2) is the squared correlation between the observed outcomes and the predicted values.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Seed | RMSE | MSE | MAE | R2 |
| 1 | 0.07642950 | 0.005841468 | 0.04521251 | 0.8025287 |
| 2 | 0.07267318 | 0.005281392 | 0.04484409 | 0.8023881 |
| 3 | 0.05788494 | 0.003350667 | 0.04528229 | 0.8031377 |
| 4 | 0.06427964 | 0.004131872 | 0.04466928 | 0.8028060 |
| 5 | 0.07050802 | 0.004971381 | 0.04500561 | 0.8027608 |
| 6 | 0.06645319 | 0.004416026 | 0.04499423 | 0.8031056 |
| 7 | 0.05580908 | 0.003114653 | 0.04529706 | 0.8029736 |
| 8 | 0.04943028 | 0.002443352 | 0.04526987 | 0.8029144 |
| 9 | 0.07920272 | 0.006273070 | 0.04475982 | 0.8018588 |
| 10 | 0.06372644 | 0.004061060 | 0.04546113 | 0.8024372 |
|  |  |  |  |  |
| Mean | 0.0656397 | 0.004388494 | 0.04507959 | 0.8026911 |

1. For Decision Tree Model: **(Paul Duong, Andrek Soto, Amir Mehra)**
   1. The model formula is “Chance of Admit ~ Serial No. + GRE Score + TOEFL Score + University Rating + SOP + LOR + CGPA + Research”
   2. Initially, we will have to create a tree with all the other variables as predictors for our response variable “Chance of Admit” which resulted in a total of 8 terminal nodes. When calling the summary of the original tree, the result we get is:  
        
      From the summary, we can see that the number of predictors actually used in the tree for their criterion is 4 (CGPA, Serial No., TOEFL.Score, and GRE.Score). Although this tree is the original tree of “Admission\_Predict” dataset, through pruning we can get a better visual representation as well as a better test error rate. Before pruning, we can calculate the test error rate of 0.844% for a comparison later. Additionally, we need to figure out what the amount of terminal nodes will be the best subset of the original tree meaning the least amount of deviance possible as well as the least amount of terminal nodes possible. If we only look at the lowest deviance, it will result in a high level of complexity. By performing pruning, we are able to get a better visual representation and better test error rate by receiving a subset of the original large tree.
   3. We divided our dataset into 80% training and 20% testing data by setting train = sample(nrow(Admission\_Predict), 0.8 \* nrow(Admission\_Predict)).
   4. As we performed the regression decision tree model on the training data, we know that a subset of the originally large tree will lead to better results, so by using cross validation and pruning, we can find the best number of terminal nodes for this dataset and calculate that mean test prediction error. When we plotted the size vs. dev of the cross validation on Admission\_Predict, we can see that the best model is between 4 and 6.   
        
      By using a for loop to find the best MSE, we can find the best number of nodes the pruned tree will need. After performing the for loop, the results are:  
        
      We can note that the best MSE is the pruned tree with 4 terminal nodes. The mean test prediction error is 1.016%.
   5. 10 iterations of Steps III and IV above (Not setting seed to keep all data the different):

|  |  |
| --- | --- |
| Iteration Counter | MSE (Mean Test Prediction Error Rate) |
| 1 (set.seed(1)) | 1.0159730% (Best = 4 terminal nodes) |
| 2 (set.seed(2)) | 0.7597011% (Best = 5 terminal nodes) |
| 3 (set.seed(3)) | 0.5282337% (Best = 4 terminal nodes) |
| 4 (set.seed(4)) | 0.6486249% (Best = 4 terminal nodes) |
| 5 (set.seed(5)) | 0.7311297% (Best = 4 terminal nodes) |
| 6 (set.seed(6)) | 0.6019986% (Best = 4 terminal nodes) |
| 7 (set.seed(7)) | 0.5586182% (Best = 6 terminal nodes) |
| 8 (set.seed(8)) | 0.5145212% (Best = 4 terminal nodes) |
| 9 (set.seed(9)) | 0.8179073% (Best = 4 terminal nodes) |
| 10 (set.seed(10)) | 0.4850815% (Best = 4 terminal nodes) |

After rerunning Steps III and IV, we can note that the majority of the tree’s best number of nodes for the pruned tree is 4, with the mean test prediction error of the model being 0.6662%.

1. Prediction Comparison and Concluding Remarks **(Amir Mehra, Andreck Soto)**
   1. For the linear model the different mean performance metrics were respectively the RMS which was equal to 0.0656397, the MSE which was equal to 0.004388494, the MAE which was equal to 0.04507959, and the R2 which was equal to 0.8026911. Additionally for the Decision Tree Model, the majority of the tree’s best number of nodes for the pruned tree is 4, with the mean test prediction error of the model being 0.6662%. When comparing the prediction test error from both the Decision Tree and Linear Models, the Linear Model has a slightly lower prediction test error than the Decision Tree Model. The Decision Tree Model has a prediction test error of 0.6662% while the Linear Model has a prediction test error of 0.4388%.
   2. Overall the linear model worked better because it maintained a stronger linear relationship. Based on the linear model we can see that the coefficient of determination was equal to 0.8026911. Hence, the predictors have strong correlation and lie close to the linear regression line and the response variable will have a more accurate result because of this.