

BAIT 509 Final Assignment

Anh Nguyen, Suguna Menon, Raymond Hu

1. Background and Motivation

1.1 Business Question

The business problem of our client, Education Consultant Agency (ECA), is to predict how likely a student will be admitted into a desired university. This is useful for ECA's clients to shortlist their targeted universities with their profiles. In addition, with the model, ECA's clients can know which areas are important to admission so that they should focus on improving. ECA is also interested in charging a premium for high-risk clients - those with lower chance of admission.

1.2 Dataset Description

The dataset is obtained from <https://www.kaggle.com/mohansacharya/graduate-admissions>.

The response variable of the dataset is 'Chance', which is an admission category of 5 levels (very low, moderately low, medium, moderately high, very high). This admission level relates to an applicant's chance of being admitted in a graduate school based on his/her profile. The applicant profile consists of seven features: CGPA, GRE Score, TOEFL Score, SOP, University rating, LOR, and Research.

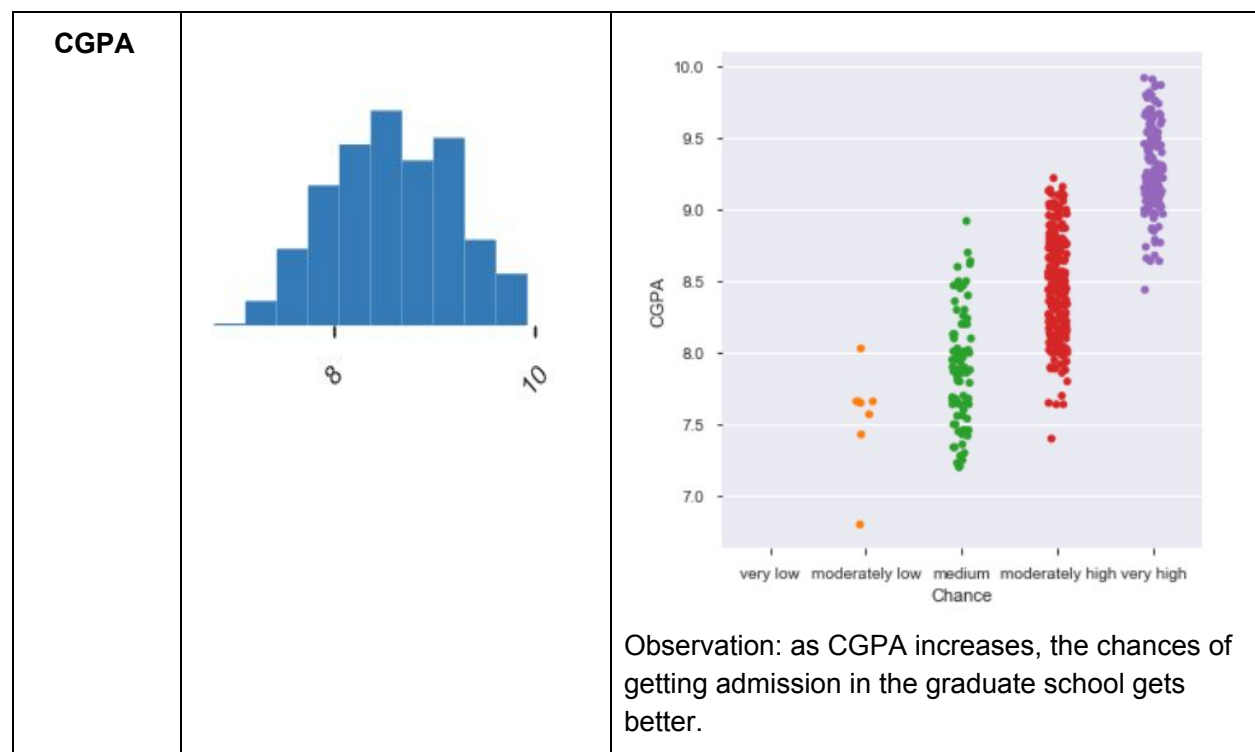
Feature	Type	Description
CGPA	Numeric	Undergraduate GPA (out of 10)
GRE Score	Numeric	GRE Scores (out of 340)
TOEFL Score	Numeric	TOEFL Score (out of 120)
University Rating	Numeric	Different tiers that universities belong to (1=strongest tier, 5=weakest tier)
SOP	Numeric	Statement of Purpose strength, as rated by agency (out of 5)
LOR	Numeric	Letter of Recommendation strength, as rated by agency (out of 5)
Research	Numeric	Research Experience (1 = have experience, 0 = no experience)

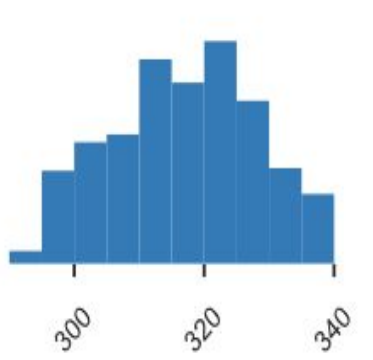
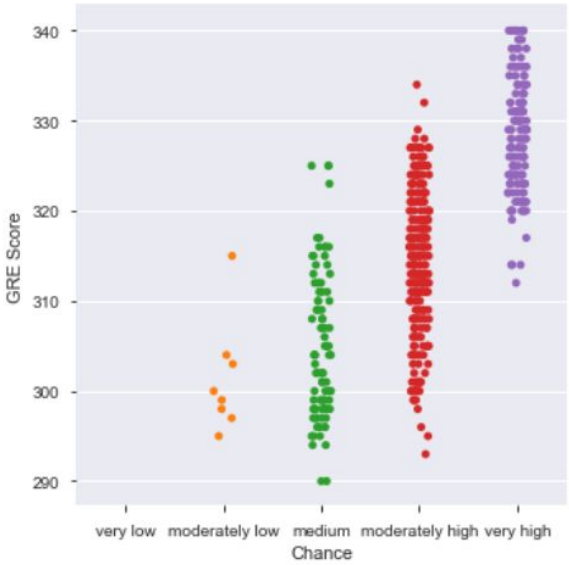
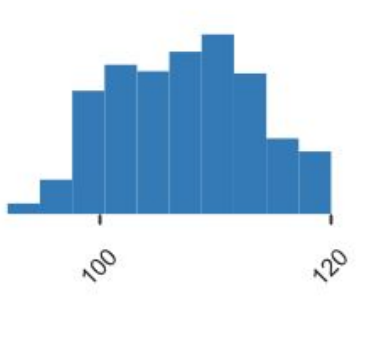
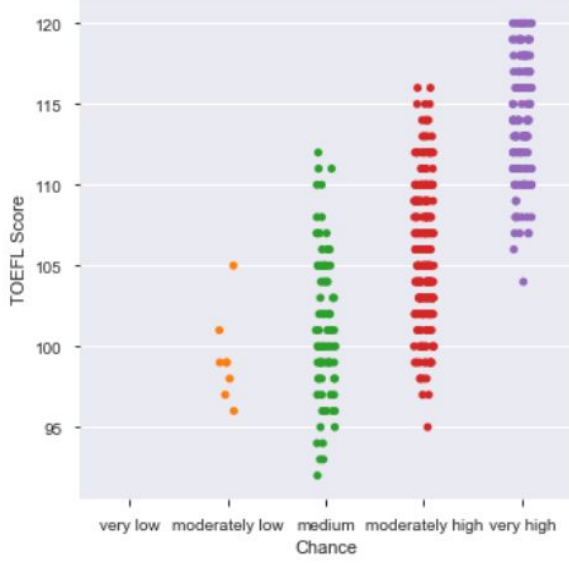
1.3 Exploratory Data Analysis

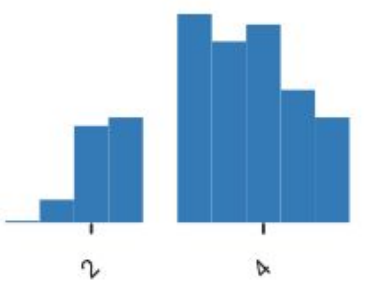
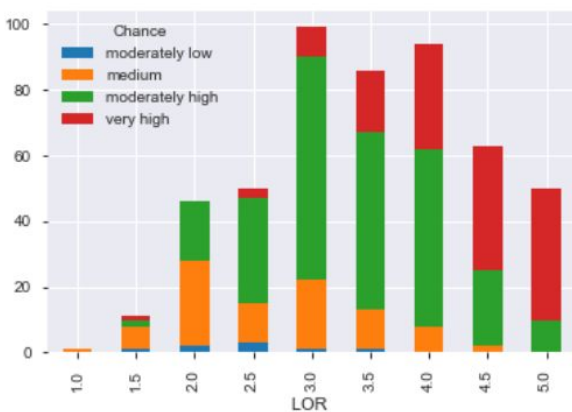
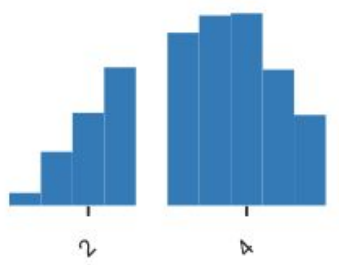
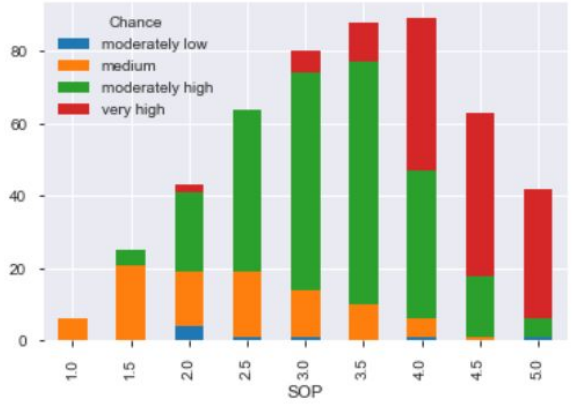
Dataset info

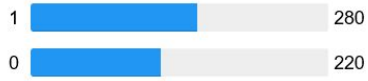
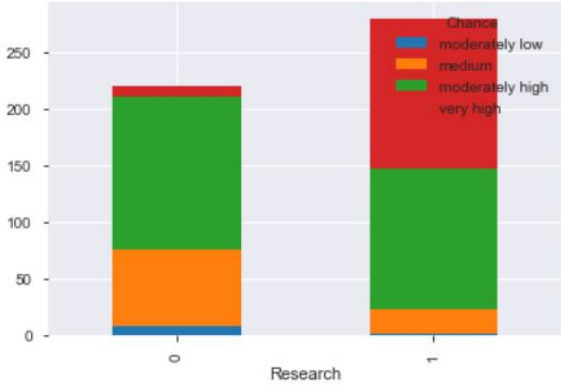
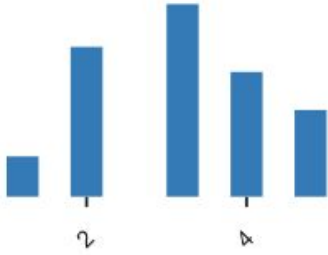
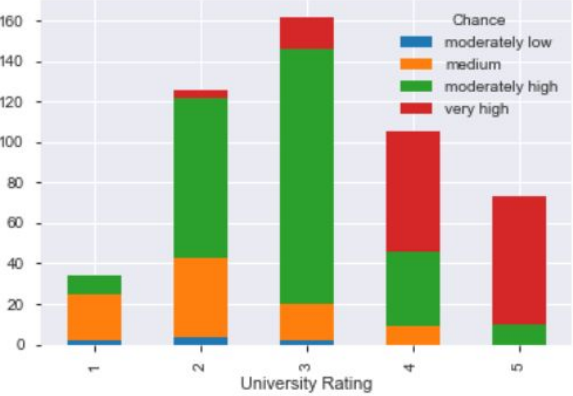
Number of variables	9
Number of observations	500
Missing cells	0 (0.0%)
Duplicate rows	0 (0.0%)
Total size in memory	32.4 KiB
Average record size in memory	66.4 B

The dataset has 9 columns, consisting of the applicant ID ('Serial No. '), the 7 features ('CGPA', 'GRE Score', 'TOEFL Score', 'SOP', 'University rating', 'LOR', 'Research'), and response variable 'Chance'. There are 500 observations in the dataset. There is no NA value.



<p>GRE Score</p>	 <p>A histogram showing the distribution of GRE scores. The x-axis is labeled with 300, 320, and 340. The distribution is roughly bell-shaped, centered around 320-325, with most scores falling between 300 and 340.</p>	 <p>A scatter plot showing GRE Score (y-axis, ranging from 290 to 340) versus Chance (x-axis, with categories: very low, moderately low, medium, moderately high, very high). The data points are colored by chance level: orange for very low, green for moderately low, red for medium, and purple for moderately high and very high. The plot shows a clear upward trend, indicating that as the GRE score increases, the chance of admission also increases.</p> <p>Observation: as GRE Score increases, the chances of getting admission in the graduate school gets better.</p>
<p>TOEFL Score</p>	 <p>A histogram showing the distribution of TOEFL scores. The x-axis is labeled with 100 and 120. The distribution is roughly bell-shaped, centered around 110-115, with most scores falling between 100 and 120.</p>	 <p>A scatter plot showing TOEFL Score (y-axis, ranging from 95 to 120) versus Chance (x-axis, with categories: very low, moderately low, medium, moderately high, very high). The data points are colored by chance level: orange for very low, green for moderately low, red for medium, and purple for moderately high and very high. The plot shows a clear upward trend, indicating that as the TOEFL score increases, the chance of admission also increases.</p> <p>Observation: With increase in TOEFL Score, the chances of getting admitted to the graduate school gets better.</p>

LOR	 <p>A histogram showing the distribution of LOR scores. The x-axis is labeled with '2' and '4'. The distribution is bimodal, with a smaller peak around 2 and a larger peak around 4.</p>	 <p>A stacked bar chart showing the count of applicants for each LOR score (1.0 to 5.0) categorized by admission chance. The legend indicates: moderately low (blue), medium (orange), moderately high (green), and very high (red).</p> <table><tr><th>LOR</th><th>moderately low</th><th>medium</th><th>moderately high</th><th>very high</th></tr><tr><td>1.0</td><td>0</td><td>1</td><td>0</td><td>0</td></tr><tr><td>1.5</td><td>0</td><td>5</td><td>2</td><td>0</td></tr><tr><td>2.0</td><td>0</td><td>25</td><td>18</td><td>0</td></tr><tr><td>2.5</td><td>2</td><td>12</td><td>32</td><td>3</td></tr><tr><td>3.0</td><td>0</td><td>22</td><td>68</td><td>8</td></tr><tr><td>3.5</td><td>0</td><td>12</td><td>53</td><td>18</td></tr><tr><td>4.0</td><td>0</td><td>8</td><td>55</td><td>32</td></tr><tr><td>4.5</td><td>0</td><td>2</td><td>22</td><td>38</td></tr><tr><td>5.0</td><td>0</td><td>0</td><td>10</td><td>40</td></tr></table> <p>Note: The data shows count of applicants Observation: If the LOR strength is between 3 to 5, chances of getting admission are better as compared to lower strength of LOR.</p>	LOR	moderately low	medium	moderately high	very high	1.0	0	1	0	0	1.5	0	5	2	0	2.0	0	25	18	0	2.5	2	12	32	3	3.0	0	22	68	8	3.5	0	12	53	18	4.0	0	8	55	32	4.5	0	2	22	38	5.0	0	0	10	40
LOR	moderately low	medium	moderately high	very high																																																
1.0	0	1	0	0																																																
1.5	0	5	2	0																																																
2.0	0	25	18	0																																																
2.5	2	12	32	3																																																
3.0	0	22	68	8																																																
3.5	0	12	53	18																																																
4.0	0	8	55	32																																																
4.5	0	2	22	38																																																
5.0	0	0	10	40																																																
SOP	 <p>A histogram showing the distribution of SOP scores. The x-axis is labeled with '2' and '4'. The distribution is bimodal, with a smaller peak around 2 and a larger peak around 4.</p>	 <p>A stacked bar chart showing the count of applicants for each SOP score (1.0 to 5.0) categorized by admission chance. The legend indicates: moderately low (blue), medium (orange), moderately high (green), and very high (red).</p> <table><tr><th>SOP</th><th>moderately low</th><th>medium</th><th>moderately high</th><th>very high</th></tr><tr><td>1.0</td><td>0</td><td>5</td><td>0</td><td>0</td></tr><tr><td>1.5</td><td>0</td><td>20</td><td>5</td><td>0</td></tr><tr><td>2.0</td><td>2</td><td>15</td><td>23</td><td>2</td></tr><tr><td>2.5</td><td>0</td><td>18</td><td>45</td><td>0</td></tr><tr><td>3.0</td><td>0</td><td>15</td><td>60</td><td>5</td></tr><tr><td>3.5</td><td>0</td><td>10</td><td>65</td><td>10</td></tr><tr><td>4.0</td><td>0</td><td>5</td><td>40</td><td>45</td></tr><tr><td>4.5</td><td>0</td><td>2</td><td>15</td><td>43</td></tr><tr><td>5.0</td><td>0</td><td>0</td><td>5</td><td>35</td></tr></table> <p>Note: The data shows count of applicants Observation: If the SOP strength is between 2.5 to 5, the chances of getting admission are better as compared to the applicants with quite low SOP strength.</p>	SOP	moderately low	medium	moderately high	very high	1.0	0	5	0	0	1.5	0	20	5	0	2.0	2	15	23	2	2.5	0	18	45	0	3.0	0	15	60	5	3.5	0	10	65	10	4.0	0	5	40	45	4.5	0	2	15	43	5.0	0	0	5	35
SOP	moderately low	medium	moderately high	very high																																																
1.0	0	5	0	0																																																
1.5	0	20	5	0																																																
2.0	2	15	23	2																																																
2.5	0	18	45	0																																																
3.0	0	15	60	5																																																
3.5	0	10	65	10																																																
4.0	0	5	40	45																																																
4.5	0	2	15	43																																																
5.0	0	0	5	35																																																

Research	 <p>1 280</p> <p>0 220</p>	 <p>Note: The data shows count of applicants Observation: If the applicant has research experience, the chances of getting admission are better as compared to the applicants with no research experience.</p>
University Rating		 <p>Note: The data shows count of applicants Observation: If the University belongs to tier 1-2, the chance of getting admission is relatively low (difficult to get in) as compared to the universities belonging to tier 3-5 (easier to get in).</p>

In general, all the histograms are almost unimodal and have no anomalies.

2. Discussion about Questions

The business questions of Education Consulting Agency can be classified into two types:

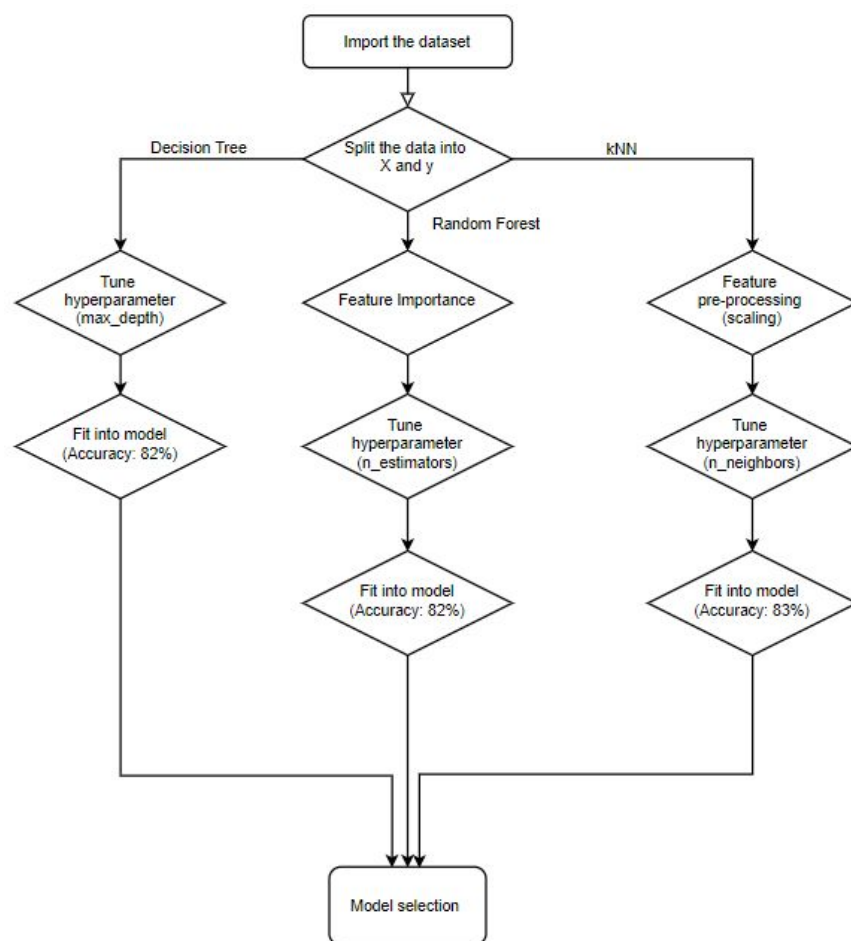
1. What is the likelihood of admission of an applicant
2. Which features have the biggest impact on the level of admission (Which features should applicants focus on the most)?

These business questions translate to the corresponding statistical questions:

1. Build a model to predict the level of admission of applicant given 7 features
 - This statistical question addresses the first business question about the likelihood of admission of an applicant
 - However, the model will fail to address the business question if applicants don't have enough 7 features or have other new features
2. Find feature importance to see which features significantly affects the level of admission
 - This statistical question addresses the second business question. The feature importance ranks the importance of feature that the model used in prediction
 - However, the result is generalized to all schools. Different universities can put different weights in their admission criteria, so the agency should be careful when giving advice about feature importance to students.

3. Supervised Learning Workflow and Models

3.1 General workflow



3.2 Supervised models

Decision Tree Classifier

We start with the simplest classification model - Decision Tree.

1. Before implementing Decision Tree, we split the dataset into training and testing sets to ensure a good final model
2. Next, we tune the hyperparameter `max_depth` by calculating training error and cross-validation error. We select the `max_depth` where cross-validation error is the lowest. Because there is a bias-variance trade-off, choose the `max_depth` where validation error is the lowest helps the model to avoid underfitting or overfitting. We decided to choose `max_depth = 2`
3. After hyperparameter tuning, we fit the model with `DecisionTreeClassifier`, using `max_depth = 2`.
4. We then use this model to predict the level of admission for test data with the test accuracy of 82% (test error = 0.18)

Random Forest Classifier

We used the Random Forest Classifier model on the data following the steps below. The idea of using random forest is to use decision trees with low bias and high variance, and the random forest classifier will average out the results from these trees, reducing the overall variance.

Thus, we get a model with low bias and low variance.

1. Using the same set of X and y, identified the features that are important and have a significant relation in predicting the level of admission. After analysis, we observed that the CGPA is the most important feature and Research is the least important. We removed the feature 'Research' from the training data and fitted the model and observed that the test error was not affected.
2. Feature scaling is not required with tree based models.
3. While using Random Forest, the length (`max_depth`) is assumed to be 0 for the individual decision trees. However, the number of trees (`n_estimators`) to be used is the hyper-parameter which has to be tuned and selected.
For tuning the hyper-parameter, we used cross validation to choose the best model as well as get the optimized value for `n_estimators` that gives the least validation error. As per our analysis, this value came out to be 150.
4. Using `n_estimators` as 150 and `max_depth` as 'None', we fitted the model again using the training dataset. We then used this model to predict the level of admission for test data with an accuracy of 82% (test error = 0.18).

KNeighborsClassifier

A `KNeighbors` (shown below as 'kNN') classifier is constructed in the model selection section.

The idea that a kNN model is used is based upon the idea that students that share a lot of similar attributes are more likely getting the same chance of admission in the graduate schools.

1. Before fitting the kNN model, the training and testing data is scaled and fitted, by scaling the data the model can predict more accurately.
2. A 5-fold cross validation method is used to inspect model scores and thus to determine the optimal hyperparameter: n_neighbors used in the model.
3. A line plot is plotted, with the x-axis as the n_neighbors and the y-axis the error scores, we obtained the training and cross-validation scores of the model. Our kNN model returns a lowest cross-validation error at n_neighbors of 14.
4. The test accuracy of our kNN model is 83% (test error = 0.17).

3.3 Model Selection

Overall, we see that all three models point in the same direction. They predict very similar results, as proved by similar accuracy rates (82%, 83%) and the test data predictions below:

	Chance	Tree_prediction	rf_prediction	kNN_prediction
Serial No.				
305	moderately high	moderately high	moderately high	moderately high
341	moderately high	moderately high	moderately high	moderately high
48	very high	very high	very high	very high
68	medium	moderately high	moderately high	moderately high
480	moderately high	moderately high	very high	very high
486	moderately high	moderately high	moderately high	moderately high
311	moderately high	moderately high	moderately high	moderately high
32	moderately high	moderately high	moderately high	moderately high
250	moderately high	moderately high	moderately high	moderately high
91	moderately high	medium	moderately high	moderately high
323	moderately high	moderately high	moderately high	moderately high
169	moderately high	medium	medium	moderately high
120	moderately high	moderately high	moderately high	moderately high
67	moderately high	very high	moderately high	moderately high
306	moderately high	moderately high	moderately high	moderately high
190	very high	very high	very high	very high
435	moderately high	moderately high	moderately high	moderately high

Table showing test data's original results, decision tree predictions, random forest predictions, and kNN predictions

For prediction purposes, we choose to use the *kNN model*. kNN model has an accuracy rate of 83%, outperforming Decision Tree model and Random Forest model (both with accuracy rate 82%) by 1%.

To evaluate significance of features, we decide to obtain feature importance from the *Random Forest model* for the following reasons:

- Feature importance is not defined for the kNN Classification algorithm. The process of obtaining feature importance for kNN is difficult. Random Forest gives nearly similar results as kNN, and has built-in feature importance attribute.

- The Decision Tree model has an optimal max_depth of 2, which means it only splits on 2 features. This means we can only see two significant values of feature importance, which is not useful for interpreting and ranking feature importances.

4. Communication of results and advice

The model that we constructed for the Education Consultant Agency is a great tool for them to increase profitability and improve service efficiency.

For any new incoming student, as long as we have the student's application information, our model can predict the level of success rate on whether the student can get admitted by graduate schools. This is useful for us to differentiate our service to students with different levels of admission success rate.

Our model has a prediction accuracy of 83%. This is a fairly competent rate. We can continue to improve the model with more data.

Our model also gives insight on what are the most important factors to influence admission rate. From our feature importance table, we have concluded that the most important 3 factors for graduate school admission are: Cumulative GPA, GRE scores and TOEFL scores.

Table of Feature Importance

Feature	Importance
CGPA	0.367111
GRE Score	0.186069
TOEFL Score	0.146916
SOP	0.112951
University Rating	0.089964
LOR	0.073554
Research	0.023435

Moreover, our model can be useful if bundled up as a mobile application with a friendlier user interface. The agency can enter each feature's score and get the prediction result for the applicant. The simplicity of the app means that any staff in the agency can use it. Then, the education agency can distribute the app to their clients, or keep it as one of their competitive advantages (ability to predict accurate chance of admission for applicants).

5. Reference

<https://stackoverflow.com/questions/50319614/count-plot-with-stacked-bars-per-hue>

BAIT 509 Lecture Materials by Tomas Beuzen