BAIT 509 FINAL PROJECT

Graduate Student Admissions

Submitted by:

Kinjal Majumdar - 60302429 Prasoon Mehta - 91346528

Abhishek Ramesh - 93339737

1. Background and Motivation

Every year thousands of students across the globe apply to graduate academic programs across North America. Viewed as a corner stone of education and professional development, students across the globe assiduously build application portfolios to optimize their chances of admission at graduate programs of their choosing. Here, by classifying the graduate student admission data, we devise a model that can accurately predict a student's admission decision based on their GRE score, TOEFL score, statement of Purpose, letters of recommendations, undergraduate CGPA and published research. In order to do so we will utilize two supervised learning techniques, logistic regression and support vector classification/machines on a training and test set of data, to predict subsequent admission outcomes.

1.1 The Business Question

What is a candidate's admission decision from a particular graduate-level university program?

1.2 Statistical Question

How does a candidate's GRE score, TOEFL score, statement of purpose, letters of recommendation and undergraduate CGPA influence their admit decisions from a particular graduate-level university program?

1.3 Merits of Building the Model

Given the volume and variety of applications candidates face when applying, we wanted to give them a means by which they could focus their efforts, and minimize their application spend. We wanted to build a model that would help students gauge their admit decisions at a particular graduate program, thereby allowing them to focus their application efforts accordingly.

1.4 General Shortcomings

i. This approach does not consider the historical acceptance rate of a university while evaluating a candidate's admission decision. The student's profile would only be one half of the data. The other half would be how the university screens and evaluates. A historical acceptance rate may be a proxy metric for university's application scrutiny level.
ii. This approach also does not consider the historical acceptance rate at each university. Since each university emphasizes on the diversity of each cohort, creating a prediction model that would triangulate the acceptance rate with candidate demographics and the expected diversity rate in the rate would help better predict an admission outcome for an aspiring student.

2. Exploratory Data Analysis

2.1.1 The dataset

We look to utilize the graduate admissions dataset (source: "https://www.kaggle.com/mohansacharya/graduate-admissions") for the purpose of our analyses.

2.1.2 Data Cleaning

To assist prediction, we ran a full-check on our dataset to identify any potential anomalies (null values, incomplete values etc.) and concluded that the given dataset was highly structured and did not comprise of any discrepancies, thereby nullifying the need for data processing / cleaning.

2.2 Plots

By creating a frequency distribution trend for all variables using Histograms, we made some notable observations as follows:

i. The GRE scores concentrate around a score of 320. The lowest possible score was noted to be at 290.
ii. The majority of candidates have TOEFL scores ranging from 100-110.
iii. The majority of candidates have SOP's ranging from average to good.
iv. Candidate CGPA's (on a scale of 10) are relatively high with a mean of approximately 8.5
v. Nearly 50% of candidates do not have research publications.
vi. Nearly 40% of candidates perceive their admit chances to be greater than 75%.
vii. The majority of the predictors barring University Rating and Research follow a normal distribution.

A summary of the correlation between all of the predictors is shown below:

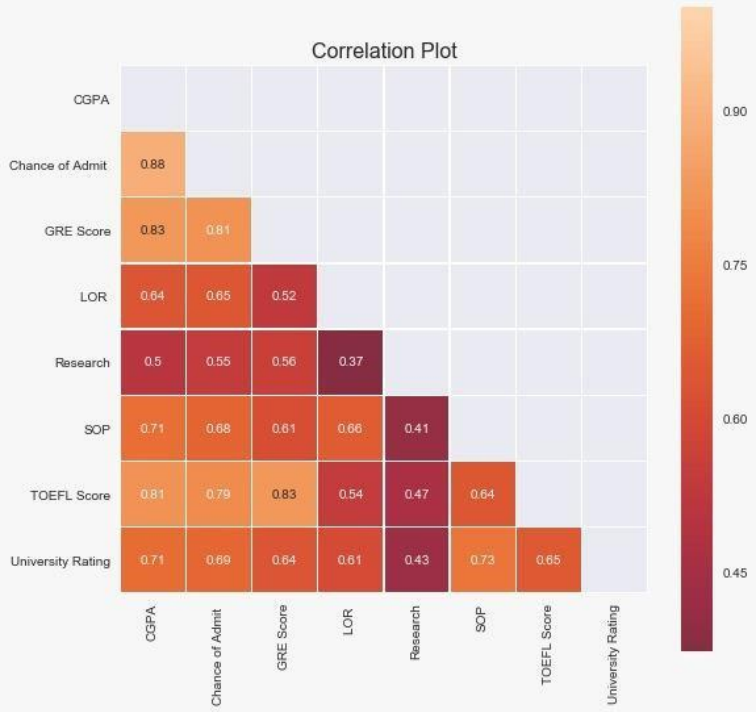

Figure 1.0: Correlation Plot of Variables

2.3 Bi-variate Analysis between Independent and Response Variables

From our bi-variate analysis we established that the independent variables influence the chances of admit in an expected fashion. The influences are elucidated below:

i. A higher GRE and TOEFL score corresponds to a higher chance of receiving admission.
ii. The higher the candidate's undergraduate university is rated the better the perceived chances of receiving admission.
iii. The higher the candidate's Statement of Purpose rating the better the perceived chances of receiving admission.

iv. The higher the candidate's Letter of Recommendation rating the better the perceived chances of receiving admission.
v. Undergraduate CGPA has the strongest influence on the chances of perceived chance of receiving admission, with a higher CGPA indicating a higher chance of perceived admit.
vi. Research experience has a weak association with perceived admit.

2.4 Correlation between Variables

We made a correlation plot, from which the following observations were noted:

i. Chance of admit is highly correlated with CGPA.
ii. GRE scores and TOEFL scores show high correlation.
iii. GRE scores are weakly correlated with both LOR & Research experience.
iv. We'd expected a high correlation between the LOR score and CGPA however there is a weak correlation

**Please refer to section 6.1 in the Appendix for Exploratory Data Analysis visualizations*

3. Model Selection

When deciding on the model to proceed with in answering our business question, we took the following considerations into account:

3.3 Quantitative choice

We decided to fit both a *Logistic Regression model, Linear Support Vector Classification* as well as a *Radial Support Vector Machines* model to the data, and subsequently proceed with the one which yields the least amount of error. If the models yield roughly the same amount of error, then we will observe the predictions and take an average of the model outputs to decide which model to go with.

3.4 Qualitative choice

In this project, we will implement and compare logistic regression and different methods of SVM classifiers: linear, and radial. To correctly apply SVM, we will preprocess the scalar to avoid any numerical dominance problems. We will also utilize the grid-search algorithm to select parameters for each SVM based on 10-fold cross-validation. Although we know that different classifiers may have different unique properties we expect the radial SVM kernel to give us the most accurate results since it is a good fit over the others on a low-dimensional feature space such as the graduate student dataset.

Given an arbitrary dataset, as the one we are handling, we typically cannot say which kernel will work best without exploring the relationships between the predictors and response- we do not know whether the problem is a linear one or a non-linear one.

Since we want to determine an output in terms of a binary 'yes' or 'no' we expect these three models to comprehensively address the task at hand, which is why we have decided to proceed with them. We can see from the EDA that CGPA, TOEFL score and GRE score show a linear association with our response of chance of admit. However, University Rating, SOP, LOR and Research all display non-linear associations with our response.

In effort to carry out the most comprehensive analysis we will first run Logistic regression and a Linear SVC model, after which we will use it to determine whether the Radial SVM is required. Our thoughts are that the Radial SVM will be the best fit for our data since we expect that the dataset will not be linearly separable.

This is because we have a limited set of data points in many dimensions. Further, SVM is good with any potential outliers as it will only use the most relevant points to find radial separation (support vectors). However, we also know that the SVM model requires tuning. In addition, the cost "C" and the use of a kernel and its parameters are critical hyper-parameters to the algorithm. Based on the loss function of the models mentioned above, we will be determine which model is appropriate. Lastly, from a review of the algorithms we considered, Radial SVM provides the best results on average.

We have listed further considerations to our model selection below:

- Since the number of features is not larger than the number of samples, the performance of the radial SVM kernel would be better than the linear kernel. This is because we may need to map into higher dimensions.
- When we map the data into lower dimensional space the radial kernel is usually better than the linear kernel owing to the fact that there are lesser features making it necessary to be mapped into nonlinear space.

Assumptions:

i. We set the cutoff values at 0.5, 0.6, 0.7, 0.8, 0.9 and iterated through them to classify.
ii. We will proceed with the best error from the cutoff values, and subsequently proceed with that cutoff.

Drawbacks of the model we expect will be best:

i. Given the complex nature of the Radial SVM model, over-fitting can become a pertinent issue.
ii. Although our assumptions are intended to reduce variance in our model, the consequence is that it increases the bias in our model.
iii. The Radial SVM kernel requires much higher computational time, and effort.

3.5 Human Choice

Since we want a model which gives an output that is easily interpretable, we decided not to go with k-NN, loess or ensemble methods. Instead, we thought that the logistic regression and SVM methods would be better at discerning easily consumable outputs to our end user.

4. Method Explanation & Results

Our objective is to classify the student data as accurately as possible to decide which receive admits from graduate programs and which students do not. To do so, let us understand how the three proposed models work.

4.1 Linear Support Vector Classifier

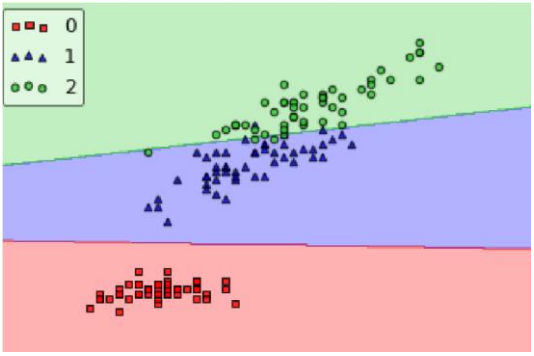

Figure 2.0: Line splitting a hypothetical dataset

As we see in Figure 1.0, this method splits up the dataset into multiple sections using a line. As the number of dimensions increase however, as is the case with our dataset, we would use a plane known as a hyperplane to segregate our data and accordingly classify the data points based on the region in which they are located. In general we choose the line/hyperplane so that the observations closest to the line/hyperplane are as far away as possible thereby minimizing the chances of a new observation being misclassified.

4.2 Radial Support Vector Machines

However, sometimes in realistic cases, as with our dataset, a more complex variant is required known as Radial Vector Machines are used to classify the data. This is useful when a dataset cannot be linearly split as shown below:

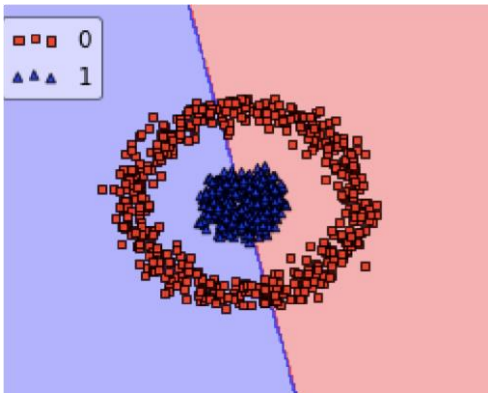

Figure 3.0: Situation where a linear classifier is not sufficient

We iterated through several cutoff values to determine the value with the least corresponding error. We observed that the least error of 0.028 was obtained for a cutoff value of 0.9.

4.3 Logistic Regression

Logistic Regression, is a mathematical model used to estimate the probability of an event occurring given some previous data. Given a set of explanatory features, which in our case are GRE score, TOEFL score, undergraduate CGPA, LOR rating, SOP rating, research and undergraduate university rating, the model determines whether a candidate will be admitted or not. The output we will receive will be either 0 or 1. In the case where student is admitted, the given value is 1. If the event does not happen, then the given value is 0.

4.4 Results

For each model we plot the cutoff values and their corresponding errors to discern which model is best suited to our needs. The cutoff scores are calculated taking each of the response parameters into account. It is interpreted as a probability.

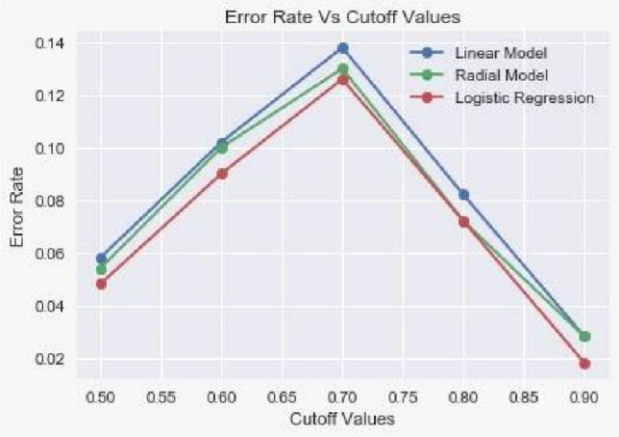

Figure 4.0: Plot of error rates obtained for each model

As we can see from the above chart, we obtain the least amount of error for all three models at a cutoff value of 0.9. Further we see that the logistic regression model gives us the least amount of error, indicating that this is the best model to go with.

We also plotted the cutoff values against the count of admits. From a university perspective, this allows an admissions to committee to decide which percentile of students they want in their program(s) and subsequently allow them to set the admission cutoff at that value. As we can see from the plot, the count of admits decreases as the cutoff increases. For tier 1 universities we would expect the cutoff to be at 0.9*

**Refer to plots 6.2.1 and 6.2.2 in the Appendix.*

Following are the corresponding optimal hyper parameters for linear SVM (Cost function – C) and for Radial SVM(Cost function-C and Gamma) for different cutoff values.

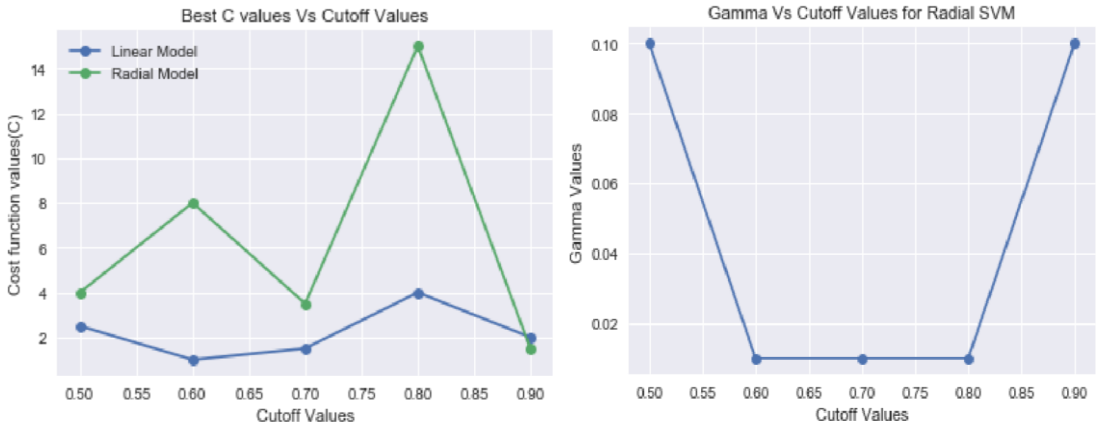

Figure 5.0: Plot of cutoff values and corresponding optimal hyperparameters – Linear & Radial SVM

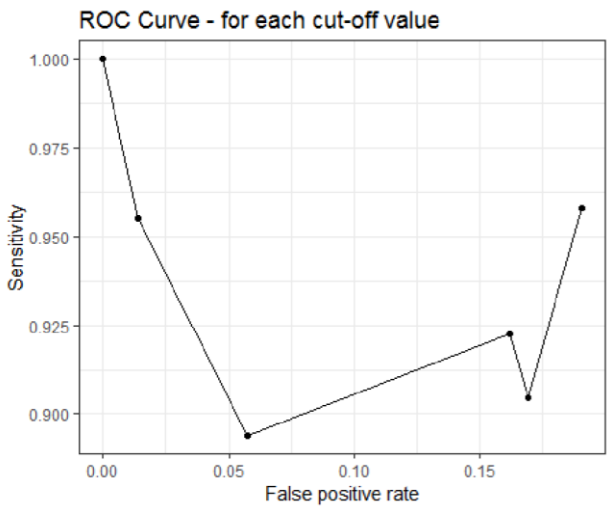

Figure 6.0: ROC curve for each cutoff value (Logistic regression model)

Note: The actual overall ROC curve is above the 'random guess' line. Figure 8.0 is a magnified version of the ROC curve which will help us visualize the cutoffs.

Each point on the ROC curve represents a sensitivity, false positive rate pair corresponding to a particular probability cutoff. A perfect ROC curve is one that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test. Our model does well and is in the upper left half near the upper left corner.

Also, in this case, we compare the ROC curve to a 'random guess line' and observe that our model performs better than the 'random guess' model which detects an accurate admit only 50% of the time.

**In depth step-by-step methodology for each of the three kernels is provided in the codes attached to this report.*

5. Recommendations

From our analysis we observed that the logistic regression model yielded the least amount of error when it was used to predict a student's chance of receiving an admit using our training dataset. To compare the different kernels, we have used 10-fold cross validation. We choose to avoid train-test split as we have limited data and cannot set aside data for validation only. In cross validation, we estimated how each model (Logistic Regression, SVM and Radial SVM) is expected to perform in general when used to make predictions on data not used during the training of the model.

We then did a comparison of each model based on their accuracy of predictions and proportion of errors, to fix on a model. In this case we would propose the use of a logistic regression model to predict a candidate's admission decision to a graduate program based on GRE score, TOEFL score, undergraduate CGPA, LOR rating, SOP rating, research and undergraduate university rating, since it is the method which gives us the highest amount of accurate predictions for the least amount of error.

The applications of this product are vast. Some of the proposals we have for this are:

i. Added feature for a graduate program study app
ii. Career counselling feature that schools and universities can offer their students
iii. Immigration and Visa services can use to profile a candidate

iv. Banks and money vendors can utilize to determine an applicant's credibility and potential
v. Admission Committees can use this to assess the type of candidates that will make up their program (based on the cutoff scores).

Implementing this as a backend software to apps that are tailored to the domains above, may give students a holistic assessment about their graduate program and professional aspirations.

Offering this service as an app in which a student can simply enter each of the parameters listed above, and receive a yes or no response as to whether they would receive a graduate program admit. To scale further, integrating datasets from colleges around the world can help create a comprehensive platform from which students can gauge their admission decision to a graduate program of their choice.

6. Appendix

6.1 Exploratory Data Analysis Plots

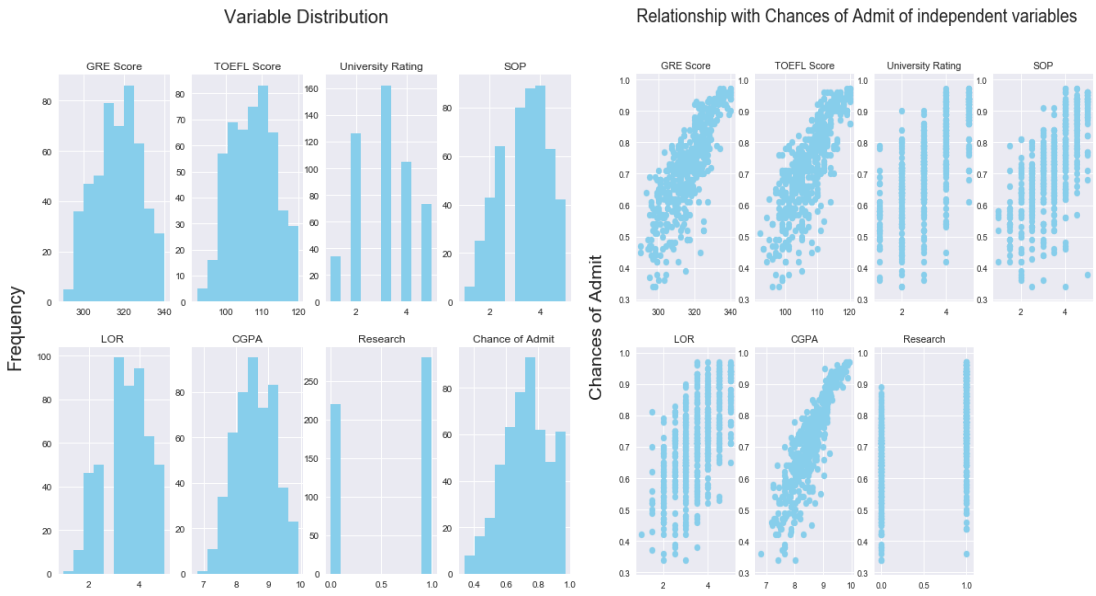

Figure 6.1.1 and 6.1.2: Distribution of features through histograms and Relationship between features and Chance.of.Admit

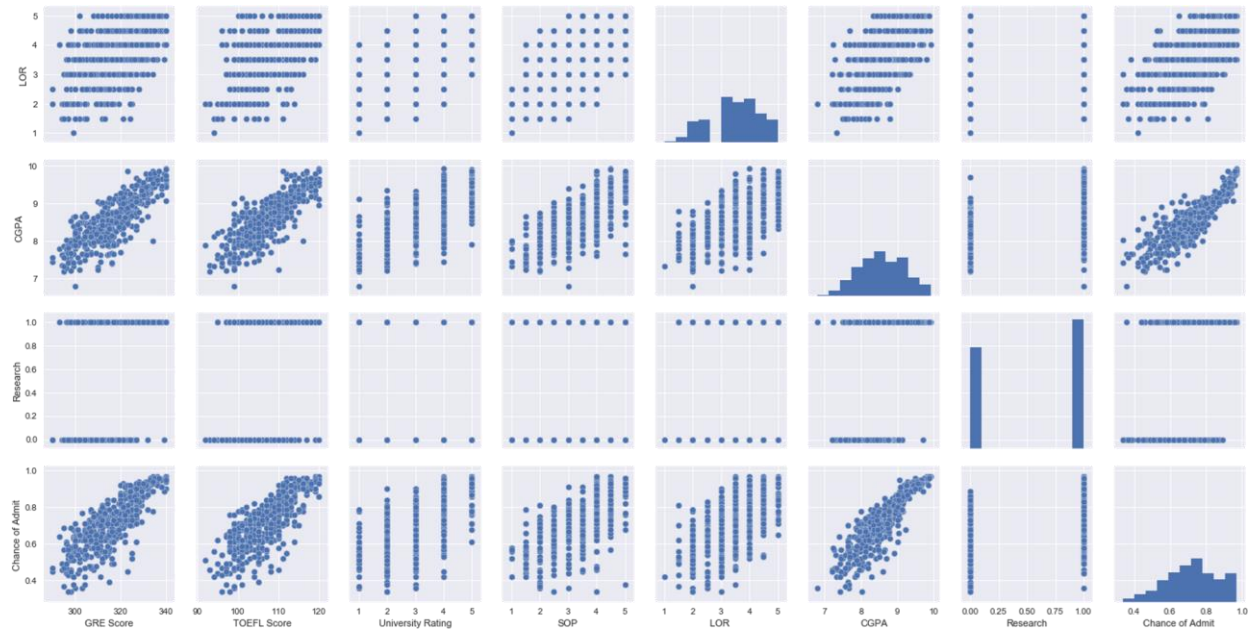

Figure 6.1.3: Plot of all variables against each other

6.2 Determining the cutoffs:

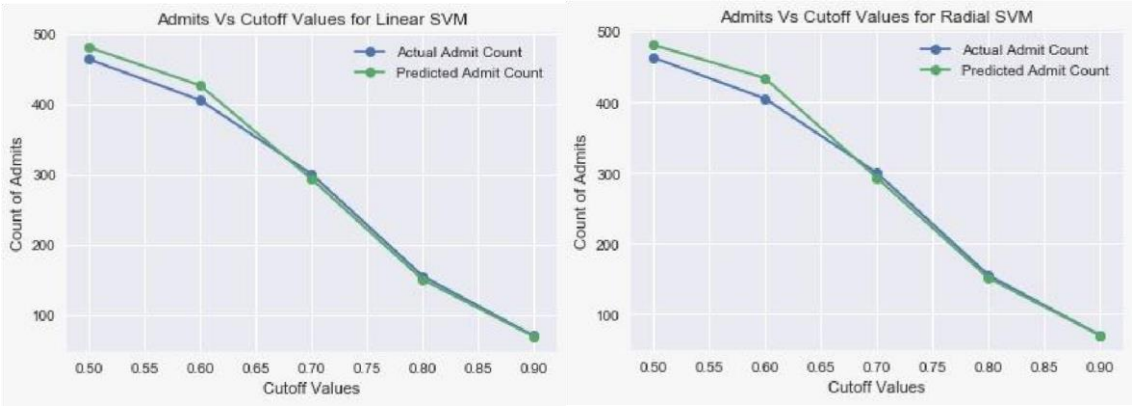

Figure 6.2.1 & 6.2.2: Plot of cutoff values and their corresponding count of admits for linear and radial SVM

These plots show us the count of admits for each cutoff value. Based on these plots we can determine for candidates and universities which graduate program cutoffs they are likely to fit.

7. References

https://vincenzocoia.github.io/BAIT509/class_meetings/cm09-svm.html

Mohan S Acharya, Asfia Armaan, Aneeta S Antony : A Comparison of Regression Models for Prediction of Graduate Admissions, IEEE International Conference on Computational Intelligence in Data Science 2019

https://datascienceplus.com/radial-kernel-support-vector-classifier/ https://www.r-bloggers.com/evaluating-logistic-regression-models/ https://www.kaggle.com/anon7r/graduate-admissions-eda-and-linear-regression

https://stackoverflow.com/questions/879173/how-to-ignore-deprecation-warnings-in-python