Predicting Graduate Admission

Hugh Morris

24/12/2020

EXECUTIVE SUMMARY

The objective of this paper is to predict the perceived chance of admission to a graduate program given a student's academic profile, that is, undergrad Grade Point Average (GPA), Graduate Record Examination (GRE), Test of English as a Foreign Language (TOEFL) score, strength of letter of recommendation (LOR) and statement of purpose (SOP), whether or not a student has research experience, as well as the ranking of the student's undergrad university.

A dataset consisting of 500 observation was split into training and testing/hold out datasets which accounted for approximately 80% and 20% of the original dataset, respectively. This was done to reduce the likelihood of overfitting the data.

Five machine learning models were trained and tuned on the training dataset using 10-fold cross validation repeated 10 times. The model that produced the lowest Root Mean Square Error (RMSE) was then chosen to predict the perceived chance of admission in the test dataset.

Exploratory data analysis was undertaken on the training dataset which suggest the presence of interaction and relatively high collinearity among the predictors, compared with the predictors and the dependent variable. This informed the model chosen to predict the chance of admission, which included linear regression with stepwise selection, as well as models that have been found to uncover interaction term between variables to inform prediction, these include random forest (bagging ensemble), support vector machine and gradient boosting (boosting ensemble). A simple linear regression model was also developed to predict the chance of admission.

The best performing model, based on the lowest RMSE was found to be Linear Regression with Stepwise Selection , which produced a RMSE of 0.0587 using the 10 fold cross validation repeated 10 times. The model selected was then applied to the test dataset and produced an RMSE of 0.0639.

Variable importance was calculated for all models using a variance based method using the individual conditional expectation curve approach. A student's GPA and GRE score were found to be the most important in determining the students perceived chance of admission into a graduate programme.

Key words: Chance of Admission, Grade Point Average, Graduate Record Examination, feature importance, Linear Regression, Linear Regression Stepwise Selection, Random Forest, Support Vector Machine, Stochastic Gradient Boosting,

SECTION I: INTRODUCTION

Globalization has opened up the opportunities for billions of persons to access goods and services outside of their countries. One of these services is education. Each year thousands of persons apply to schools outside of their country to access this service. The chances of being admitted to their desired school is based on, among other things, the academic profile of the individual. Given the time and cost to apply to these schools, it is important that potential students get an idea of their chance of getting into their desired school before they apply.

The purpose of this report is to build a model to predict the chance of entering a university graduate program based on an individual's academic profile. The paper will utilize an updated dataset (admission_predict) from (Acharya, Armaan, and Antony 2019)[1], that captures information on 500 academic profiles and their perceived chance of getting into their desired graduate programs. This dataset is applicable to persons where English is not their first language.

The rest of the paper is structured as follows:

Section II: will describe the data and undertake exploratory data analysis

Section III: will present the methodology used to build the model

Section IV: will present the results of the model

Section V: will conclude with summary of findings, limitations and further work on the topic.

SECTION II: Data Description and Exploratory Data Analysis

This section of the report seeks to give a brief description of the admission_predict dataset that will be used in the analysis, and undertake exploratory Data Analysis.

Data Description

The dataset consist of 9 variables and 500 observations:

- Serial No. - represents a unique identifier for each individual
- Graduate Record Examination (GRE) Scores (out of 340)
- Test of English as a Foreign Language (TOEFL) Scores (out of 120)
- University Rating of the applicant, which is rated from 1 to 5, with 1 being the lowest and 5 the highest. This data will be treated as categorical
- Statement of Purpose (SOP) strength, which rates the SOP from 1 to 5, with increments of 0.5. This data will be treated as numeric
- Letter of Recommendation (LOR) strength, which rates the LOR from 1 to 5, with increments of 0.5. This data will be treated as numeric
- Undergraduate/College Grade Point Average (CGPA) { out of 10 }.
- Research, which is either 0 or 1, 0 represents no research experience and 1 represents research experience. research indicates if the student has done an internship/equivalent research.
- Chance of Admit (ranging from 0 to 1)

The objective of this research paper is to predict the Chance of Admit based on the academic profile (other indicators) of each person. Therefore, not all variables will be included in the analysis (e.g. Serial No.) and some variables will be transformed to assist in analysis, for example, the University Rating and Research Experience were recoded as categorical variables.

[1] Accessed 2020-12-06: https://www.kaggle.com/mohansacharya/graduate-admissions?select=Admission_Predict_Ver1.1.csv

To guard against overfitting, the dataset was subdivided into a training dataset, which represented 80% of the dataset and a testing dataset, which represents the remaining 20% of the admission_predict dataset. The training dataset will be used for model development and parameter tuning and the testing dataset represents the final holdout dataset from which the final model will be accessed.

Exploratory Data Analysis

The remainder of this section will focus on the training dataset. The objective is to understand the data to inform model development.

Table 1 presents a distributional summary of the numerical variables. The Chance of Admit, CGPA and TOEFL score have means that are above their medians, while the other variables have means below their medians.

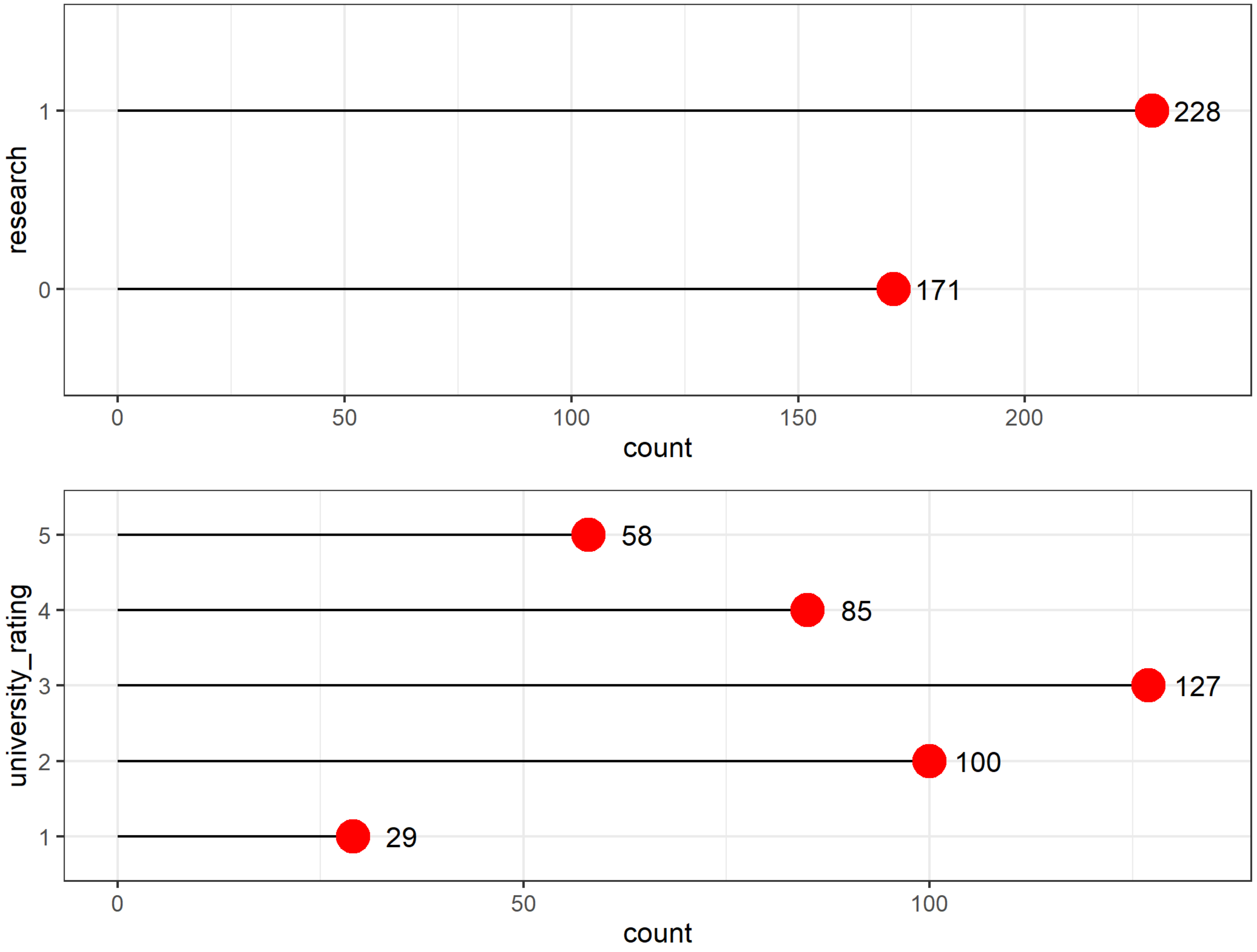
TABLE 1

Summary Statistics of Numeric Variable

Indicator	chance.of.admit	cgpa	gre_score	toefl_score	sop	lor
min	0.3400	6.800	290.0	93.0	1.000	1.00
1st Quartile	0.6300	8.120	309.0	103.0	2.500	3.00
Median	0.7200	8.570	317.0	107.0	3.500	3.50
Mean	0.7208	8.576	316.6	107.2	3.361	3.46
3rd Quartile	0.8200	9.045	325.0	112.0	4.000	4.00
Max	0.9700	9.920	340.0	120.0	5.000	5.00

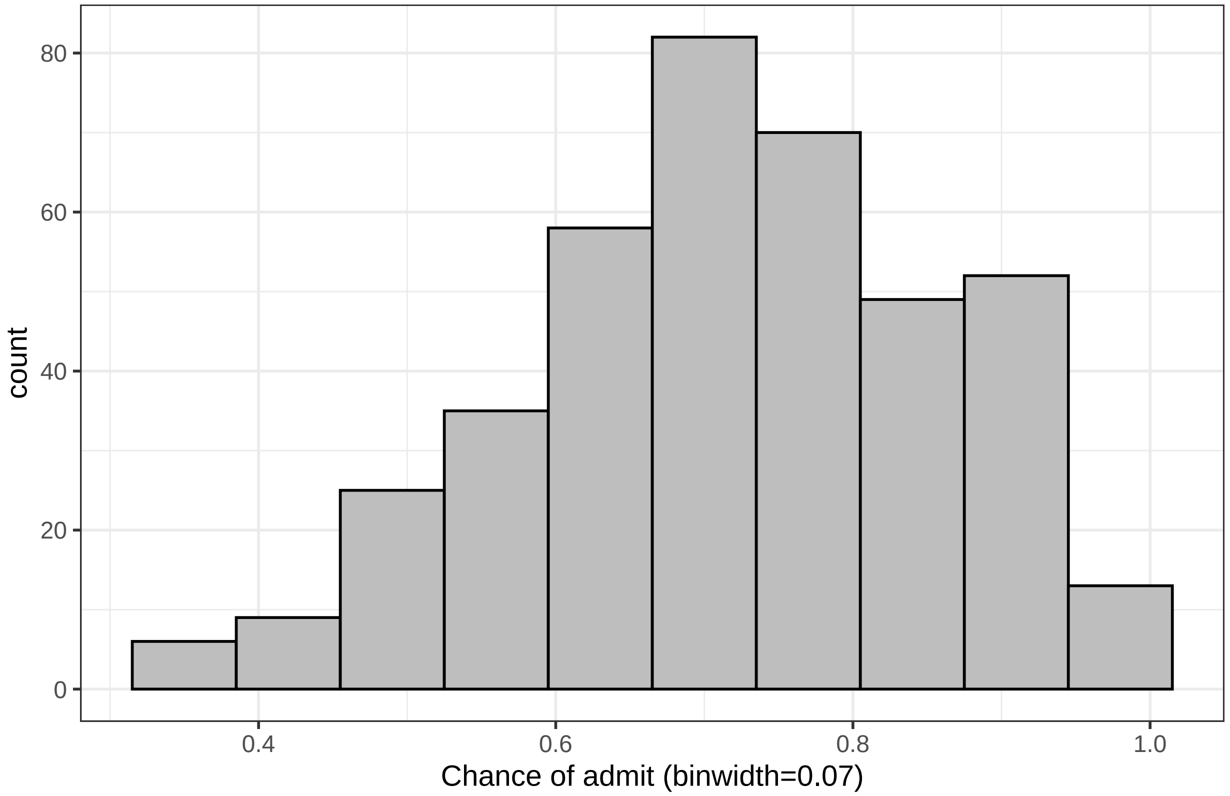
Majority of the persons in the training dataset have research experience, while majority of the persons who applied to a graduate program were from the third ranked (middle) universities (Figure 1). Research experience and the ranking of the university a student attended during undergrad are expected to influence the other academic profile of the students.

Figure 1: Frequency Distribution of Research and University Rating



Chance of admission

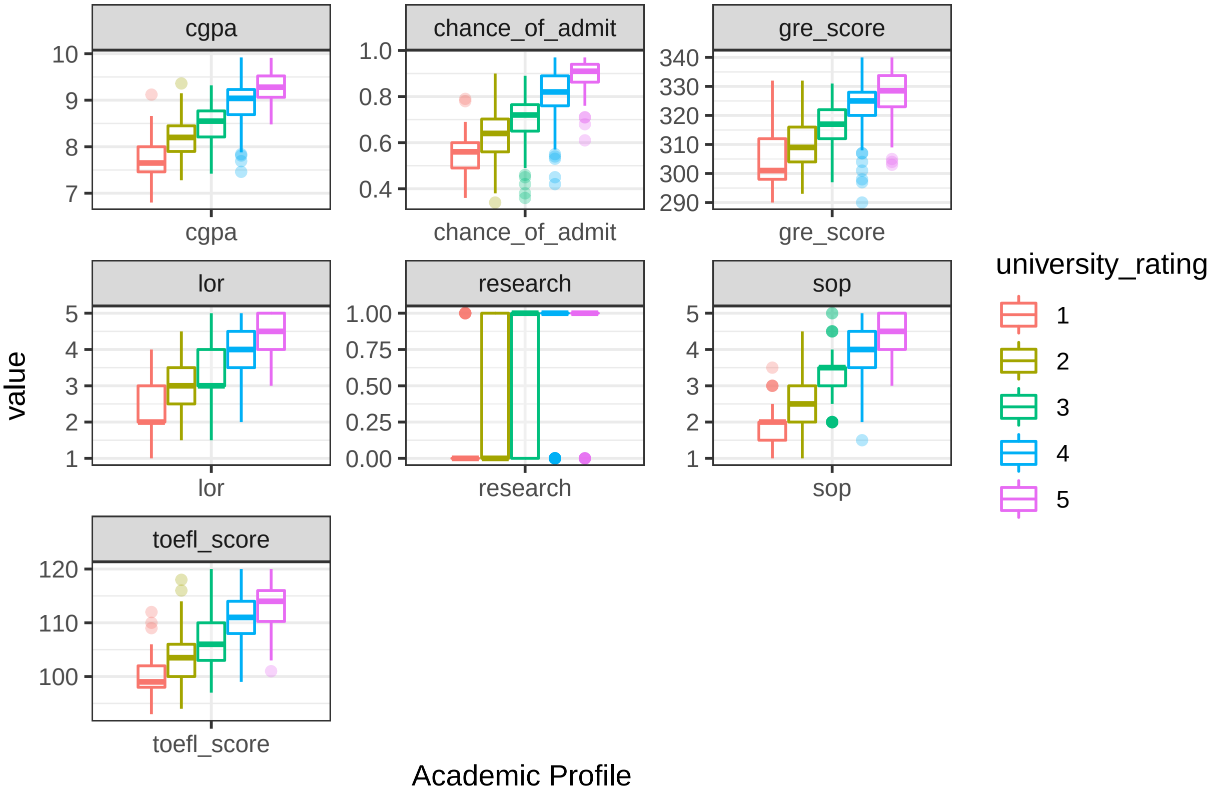
Figure 2: Histogram of Chance of Admit



Majority of the chance of admit distribution is to the right, suggesting that students are more likely to perceive that their chance of getting into a graduate program is relatively high than low.

University Ranking

Figure 3: Academic Profile Disaggregated by Applicant University Ranking

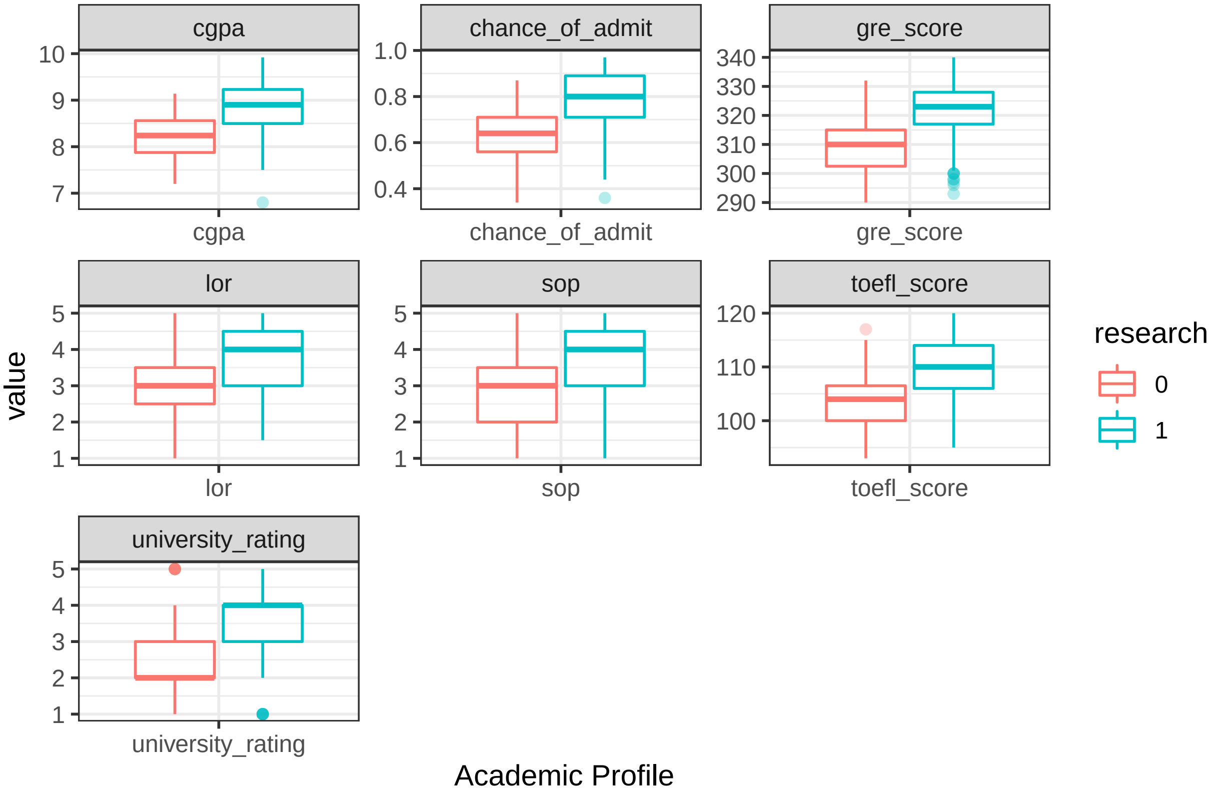


Universities ranking influences the academic profile of persons. Persons from top universities are more likely to have better academic profiles than persons from lower rank universities (Figure 3). The chance of admission increases the higher the ranking of the university.

The outliers are on the downside for the chance of admit when disaggregated by university rating, suggesting that some applicants perceive their chances of admission to a graduate program to be lower than what you expect from someone based on the rating of her/his university. This may be linked to the negative outliers in the GRE score and lack of research experience for some persons in the higher ranked universities

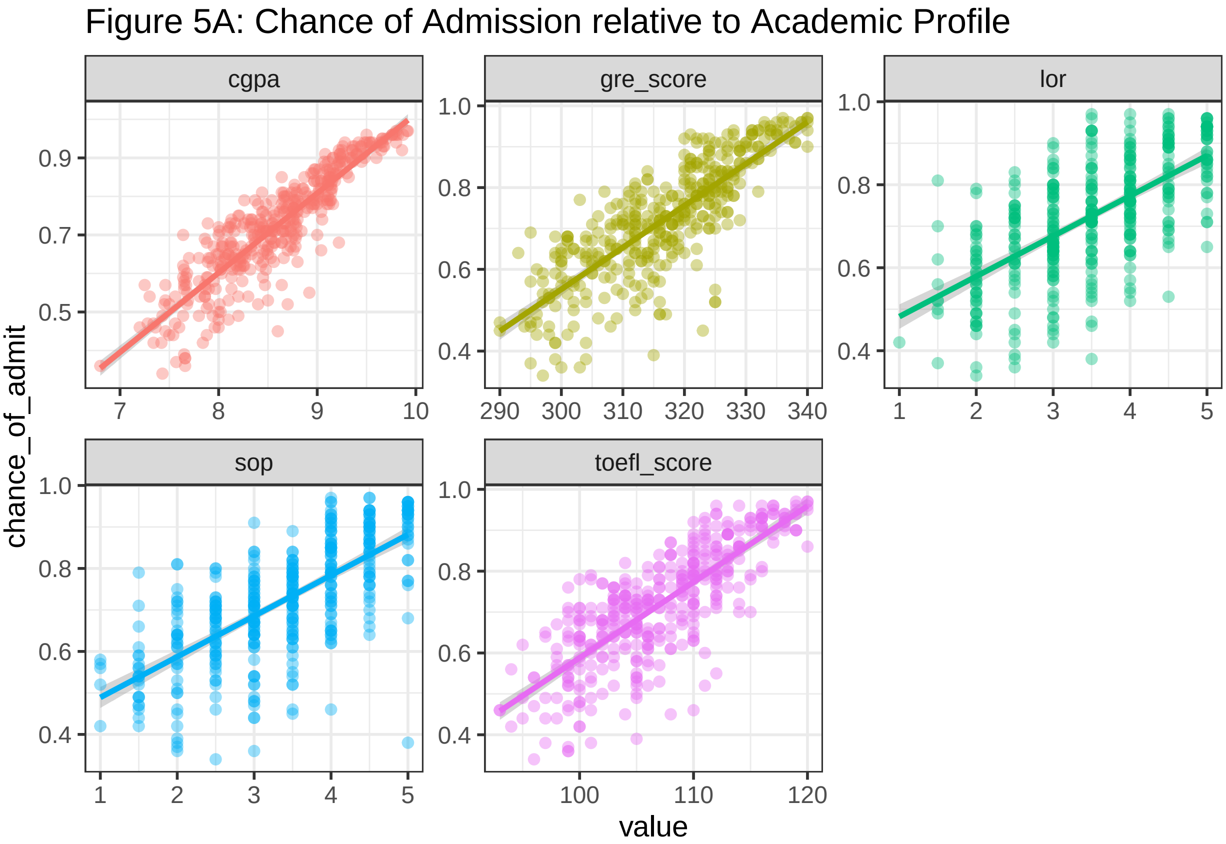
Research Experience

Figure 4: Academic Profile Disaggregated by Research Experience



Applicants with research experience are more likely to have a better academic profile than persons without research experience (Figure 4). Persons with research experience have a higher perception that they will be accepted to their desired master program (chance of admission).

Bivariate Analysis



There is a positive linear relationship between the numeric variables within the model (Figure 5A). Given the importance of the university ranking (see Figure 3) and research (see Figure 4), the numerical variables were further disaggregated by university ranking (Figure 5B) and research (Figure 5C) to identify possible interaction effects.

Figure 5B: Chance of Admission relative to Academic Profile by University Rating

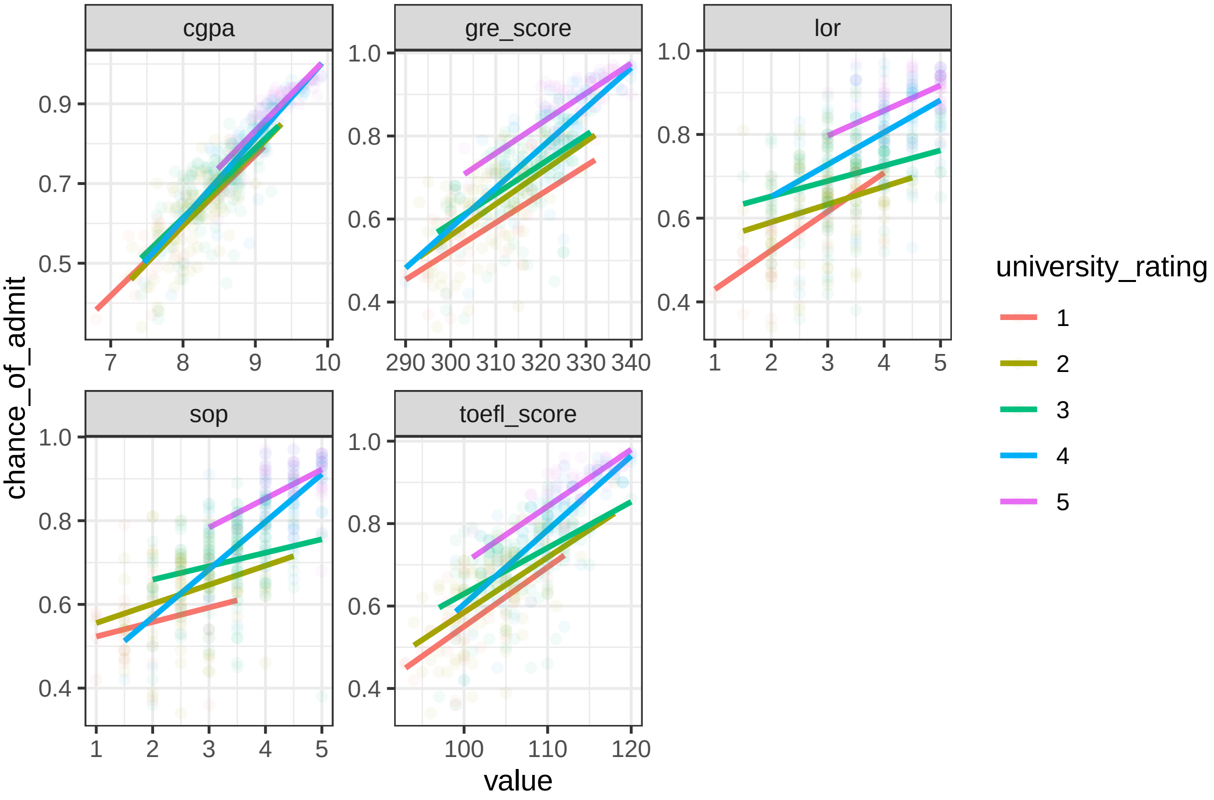
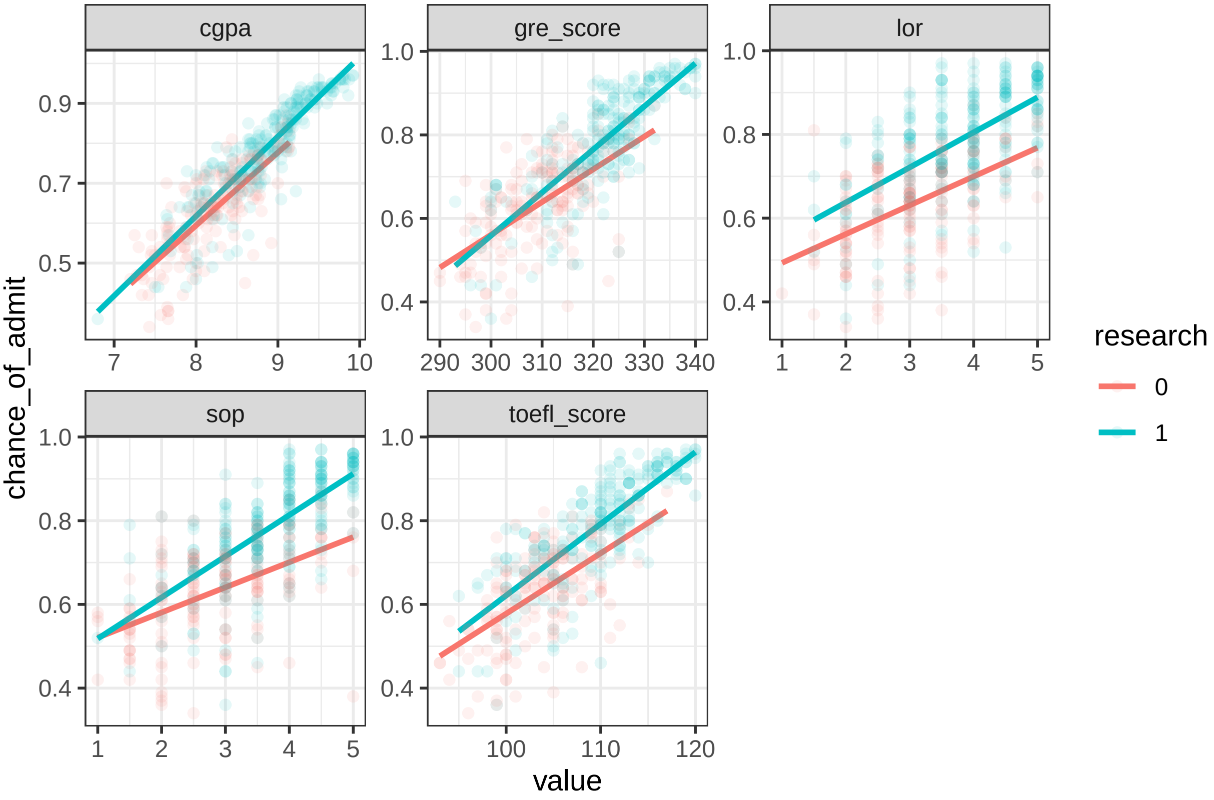
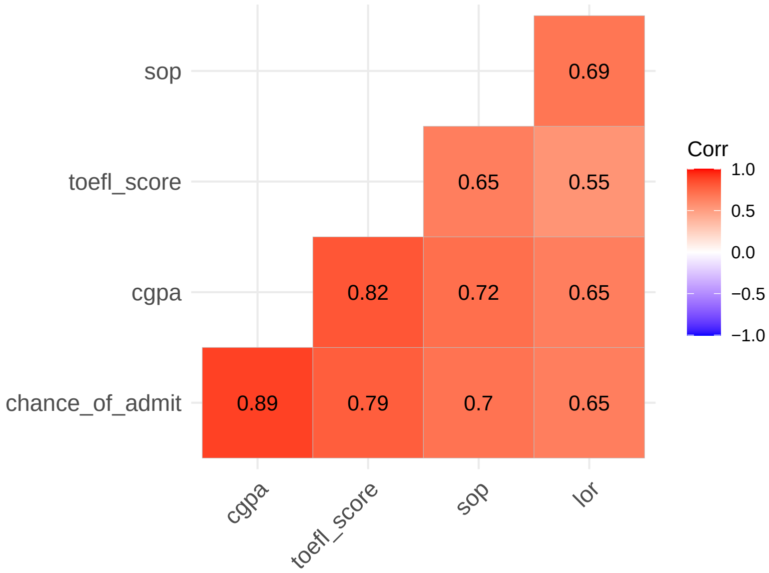


Figure 5C: Chance of Admission relative to Academic Profile by Research Expereince



The difference in the slopes, especially the crossing of the regression line in both Figure 5B and Figure 5C, suggests that there are potential interactions between the numeric variables and the categorical variables, when university rating and research experience were controlled for, respectively. These potential interaction suggests certain modeling techniques such as boosted tree-based models(Elith, Leathwick, and Hastie 2008), random forest (García-Magariños et al. 2009) and support vector machine(Chen et al. 2008), which have been found to be useful in uncovering interactions between variables, be utilized.

Figure 6: Correlation Plot



All potential numeric predictors are positively correlated with the chance of admit and with each other (Figure 6). CGPA (college GPA) has a higher correlation with GRE Score and TOEFL Score, than these variables have with the Chance of admit, suggesting the possibility of multicollinearity. One implication of multicollinearity is that some variables are redundant. This suggest a stepwise selection modeling approach may be useful.

SECTION III: METHODOLOGY

The exploratory data analysis suggest the presence of interaction variables and the potential redundancy of some of the variables. Therefore the modeling approach used consists of:

- Linear Regression with Stepwise selection model
- Random Forest Model
- Support Vector Machine Model
- Boosted Tree Based Model

These machine learning models will be compared with a linear regression model.

All models will be built and tuned using the training data set using 10-fold cross validation repeated 10 times to improve the robustness of the results. The tuning parameter that produces the lowest RMSE will be chosen to select the tuning parameters. An error band will also be produced for each RMSE using the standard deviation of the RMSE.

Feature importance will be calculated for each model using the variance-based method using the individual condition expectation curves approach to account for the possibility of interaction between the predictors(Greenwell and Boehmke 2020). The final models under each modeling approach (Boosted Tree Based Model, Support Vector Machine, etc) will be assessed against each other with the one that produces the lowest RMSE chosen.

Pre Processing:

Before the models are constructed, dummy variables are created for the two categorical variables (university ratings and research experience). Additionally, the variables are centered, that is, the average for each variable is subtracted from each observation for that variable and scaled, meaning, each variable is divided by its standard deviation. This is done for all variables in the training and test set.

SECTION IV: RESULTS

This section of the report will present the results under each machine learning model used, as well as the performance of the best performing model on the test/holdout dataset.

Linear Regression Model

The linear regression model produced a minimum RMSE of 0.0591 over the 10-fold cross validation repeated ten times (Figure 7a). The model that included the intercept produced the lowest RMSE.

Figure 7: Linear Regression Diagnostics

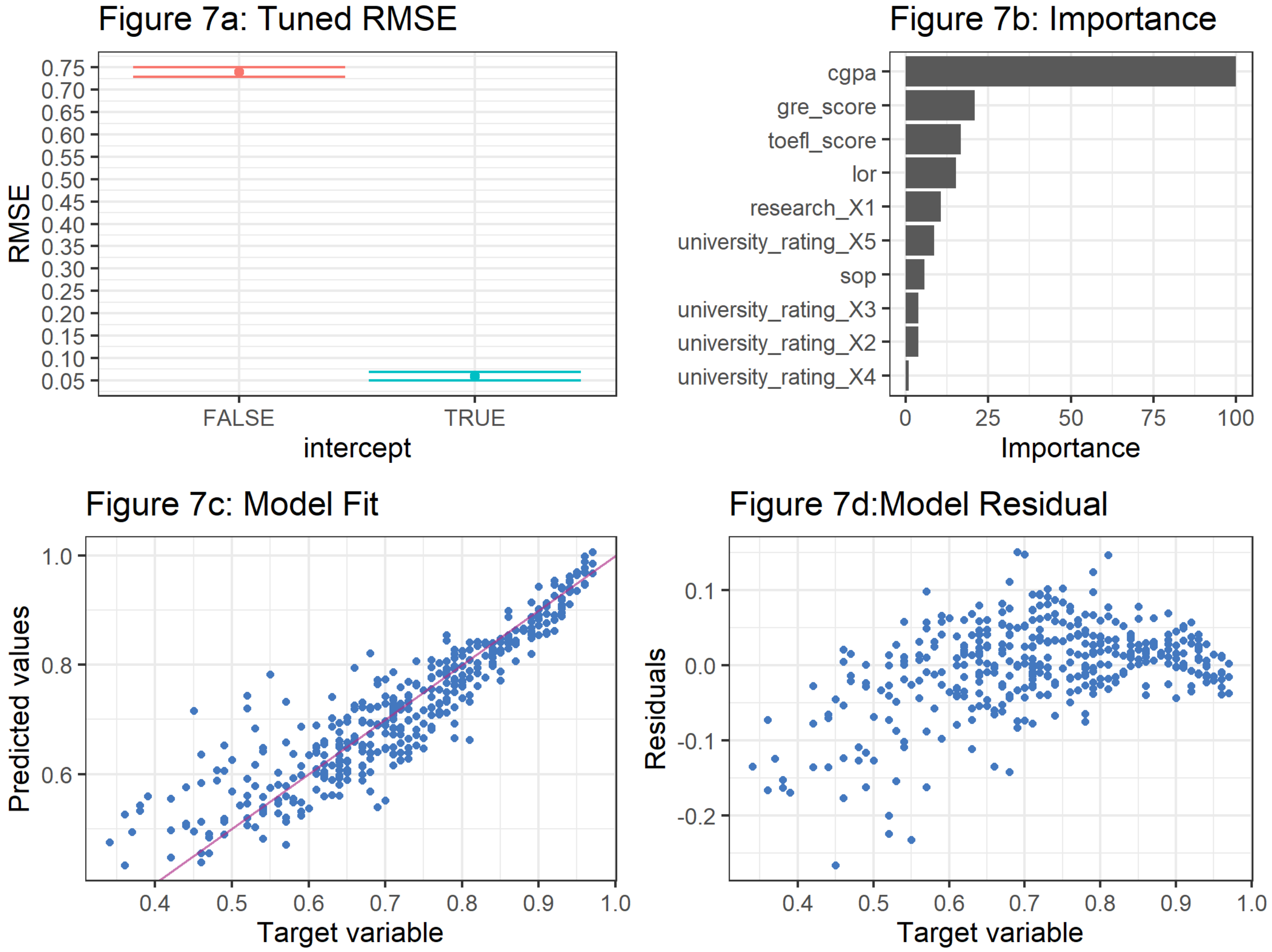


Figure 7a: Tuned RMSE

Figure 7b: Importance

Figure 7c: Model Fit

Figure 7d:Model Residual

A student's CGPA and GRE score were found to be the most important variables in predicting the perceived chance of admit (Figure 7b).

The model overestimated the perceived chance of admit when the chance of admit was relatively low (Figure 7c and Figure 7d).

Linear Regression with Stepwise Selection Model

The linear regression with stepwise selection model selected 6 of the variables to be included in the model, as well as an intercept (Table 2 and Figure 8a).

TABLE 2

Variable Selected in Linear Stepwise Model

model	Variables	Selection
model_6	Intercept	Included
model_6	gre_score	Included
model_6	toefl_score	Included
model_6	lor	Included
model_6	cgpa	Included
model_6	university_rating_X5	Included
model_6	research_X1	Included
model_6	sop	Excluded
model_6	university_rating_X2	Excluded
model_6	university_rating_X3	Excluded
model_6	university_rating_X4	Excluded

The variables that had the highest importance were CGPA and GRE scores, similar to the linear regression model (Figure 8b).

Similar to the the linear regression model, the linear regression model with stepwise selection overestimates the perceived chance of admit at a relatively small chance of admit (Figure 8c and Figure 8d)

Figure 8: Linear Stepwise Regression Results and Diagnostics

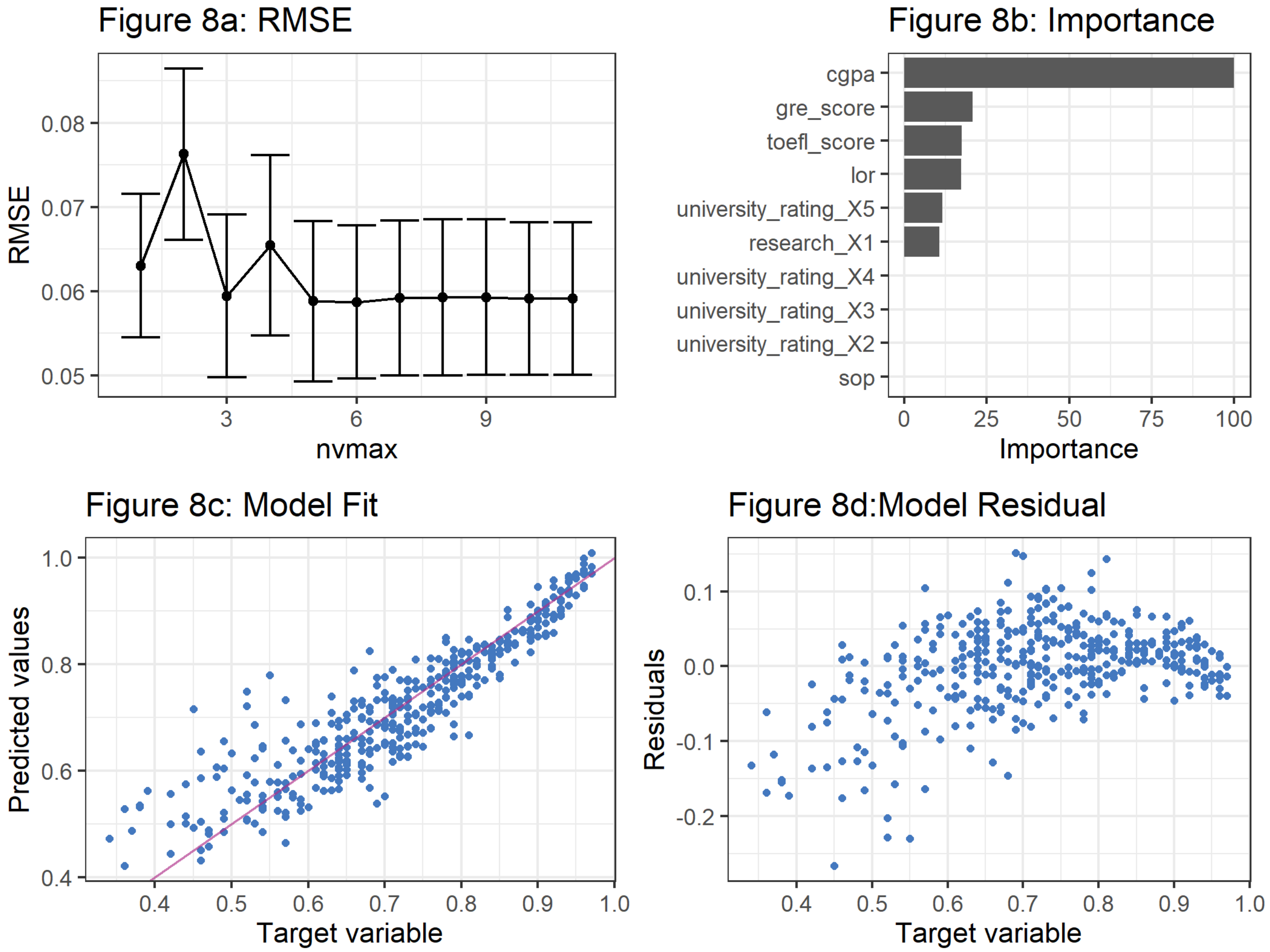


Figure 8a: RMSE

Figure 8b: Importance

Figure 8c: Model Fit

Figure 8d:Model Residual

Random Forest Model

The Random Forest model (bagging ensemble) was optimized when the number of variables that were randomly sampled at each split was 3 using the RMSE metric (Figure 9a). Similar to the previous two models, CGPA and GRE scores were found to be the most important variables in determining a person perceived chance of admit (Figure 9b).

The fit is relatively strong at higher chance of admit than at lower chance of admit (Figure 9c and Figure 9d). The random forest model is more likely to overestimate the chance of admit at lower chances of admit.

Figure 9: Random Forest Results and Diagnostics

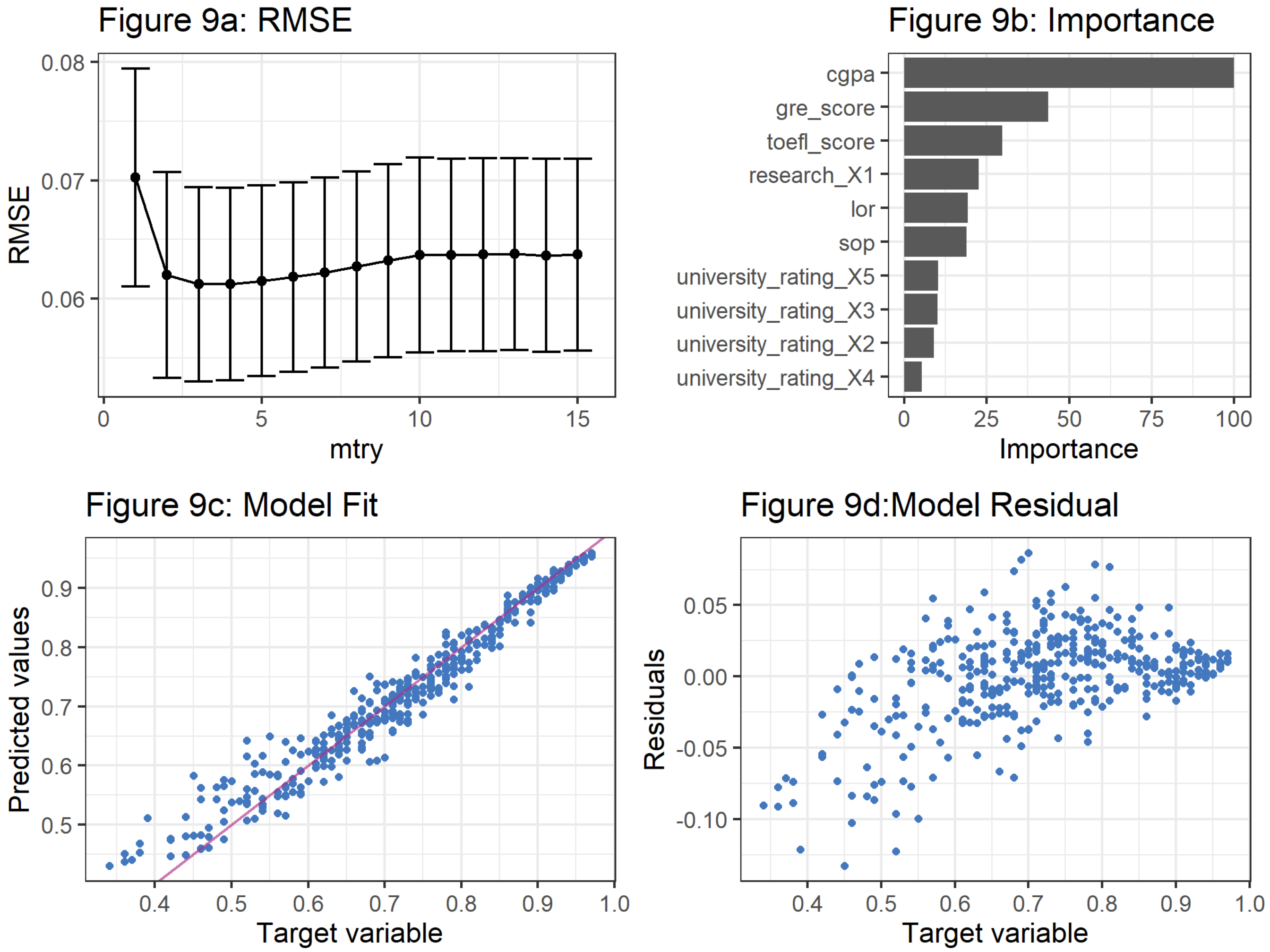


Figure 9a: RMSE

Figure 9b: Importance

Figure 9c: Model Fit

Figure 9d:Model Residual

Support Vector Machine - Support Vector Machines with Polynomial Kernel

The support vector machines (SVM) with polynomial kernel was one of the models used to predict the chance of admit. Its parameters were tuned and the model that produced the lowest RMSE chosen (Figure 10a).

The applicant's GPA and GRE Score were the most important variable that predict the perceived change of admission (Figure 10b).

Similar to the previous three models, the SVM model produced a better fit when the chance of admit was relatively high. Additionally, the model tends to overestimate at lower levels (Figure 10c and Figure 10d).

Figure 10: Support Vector Machines with Polynomial Kernel Results and Diagnostics

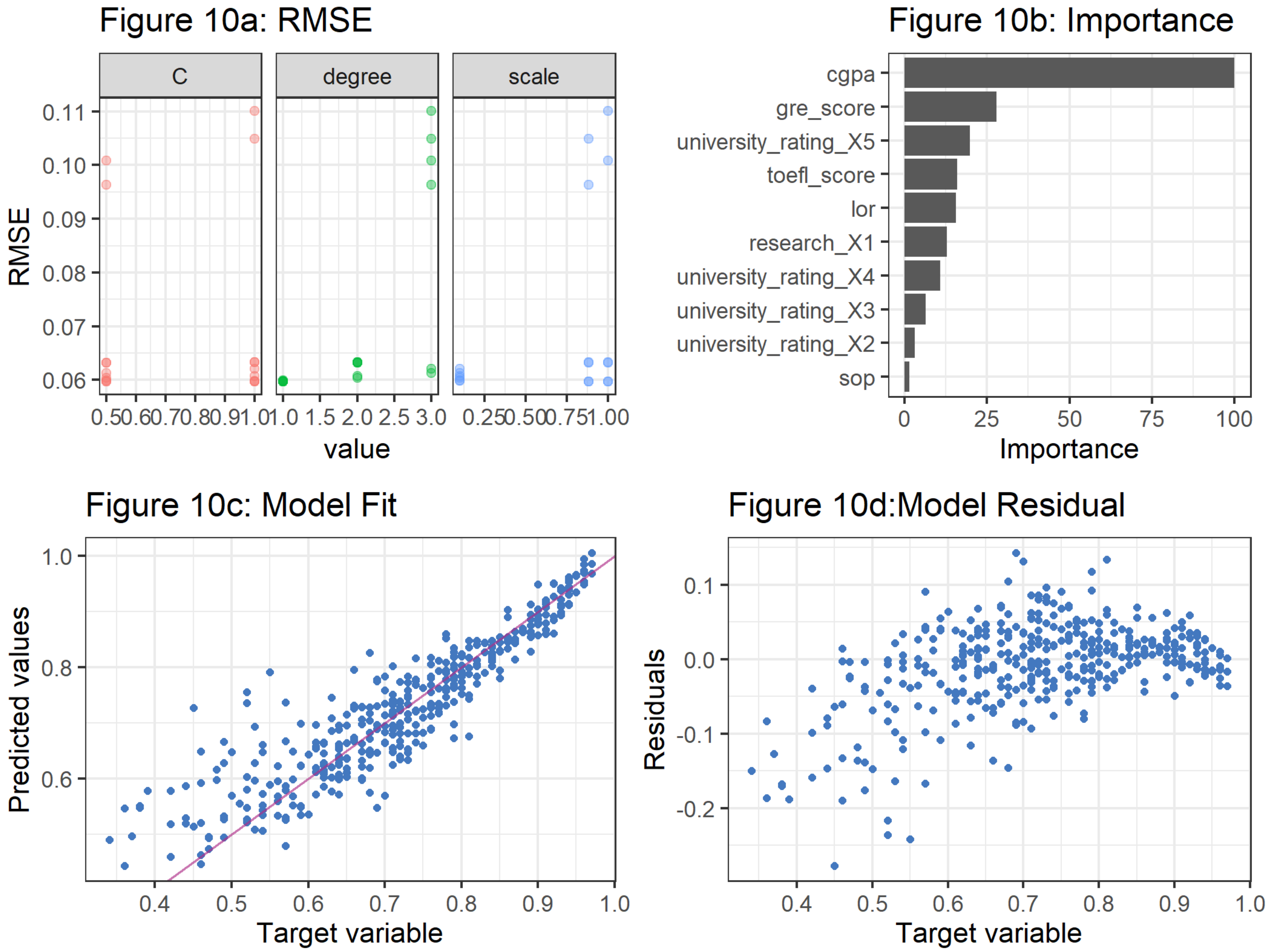


Figure 10a: RMSE

Figure 10b: Importance

Figure 10c: Model Fit

Figure 10d:Model Residual

Stochastic Gradient boosting

The Gradient boosting model (boosting ensemble) was optimized when the interaction depth was equal to 2, suggesting that there were interaction effects (Figure 11a).

Similar to the previous models the variables that were found to be most important were the applicant's GPA and GRE score (Figure 11b).

Additionally, the model's performance was comparable to the previous models, as it performed reasonably well in predicting relatively high levels of perceived chance of admit and overestimated at relatively low levels of perceived chance of admit (Figure 11c and 11d).

Figure 11: Stochastic Gradient Boosting Results and Diagnostics

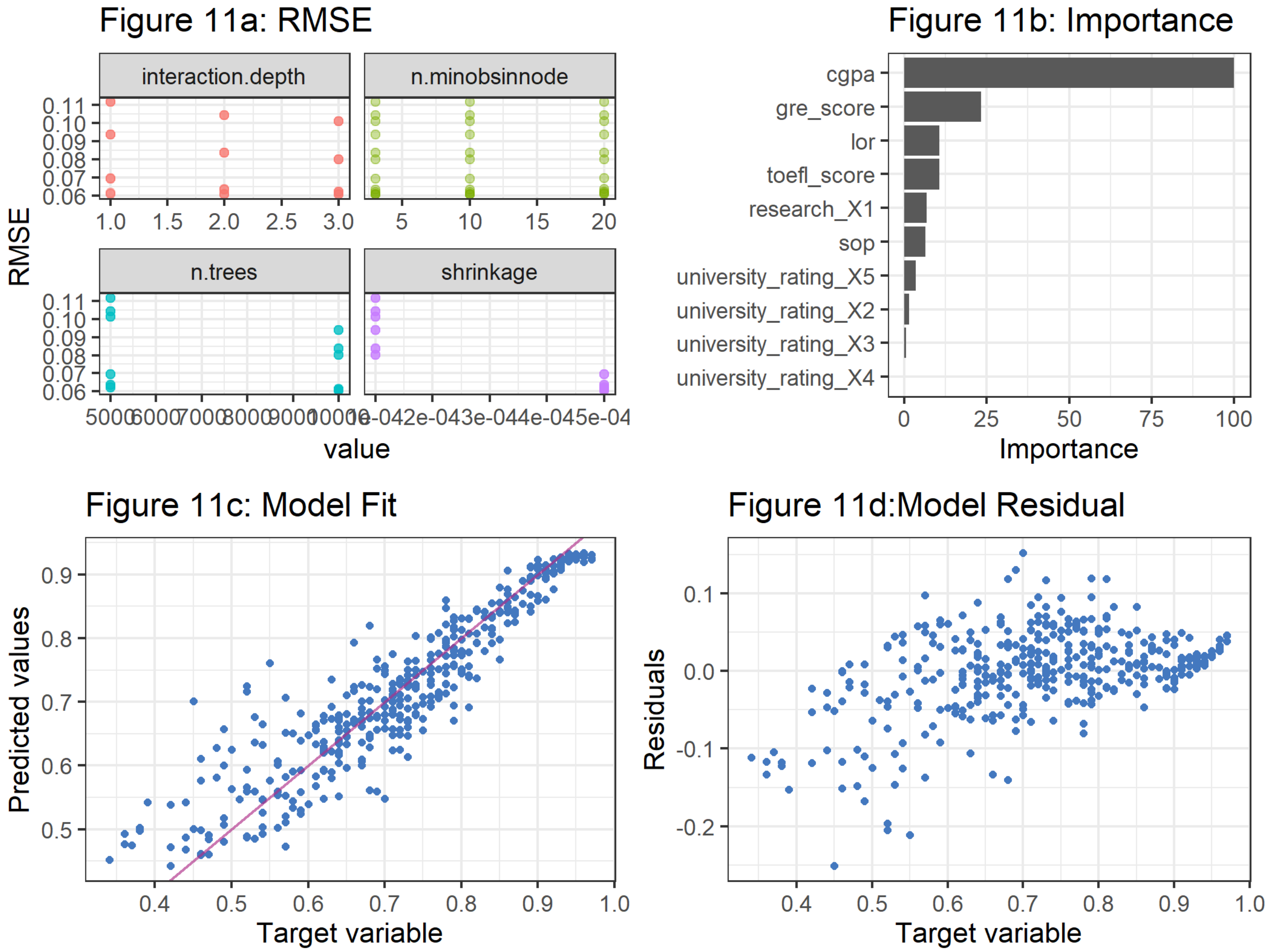


Figure 11a: RMSE

Figure 11b: Importance

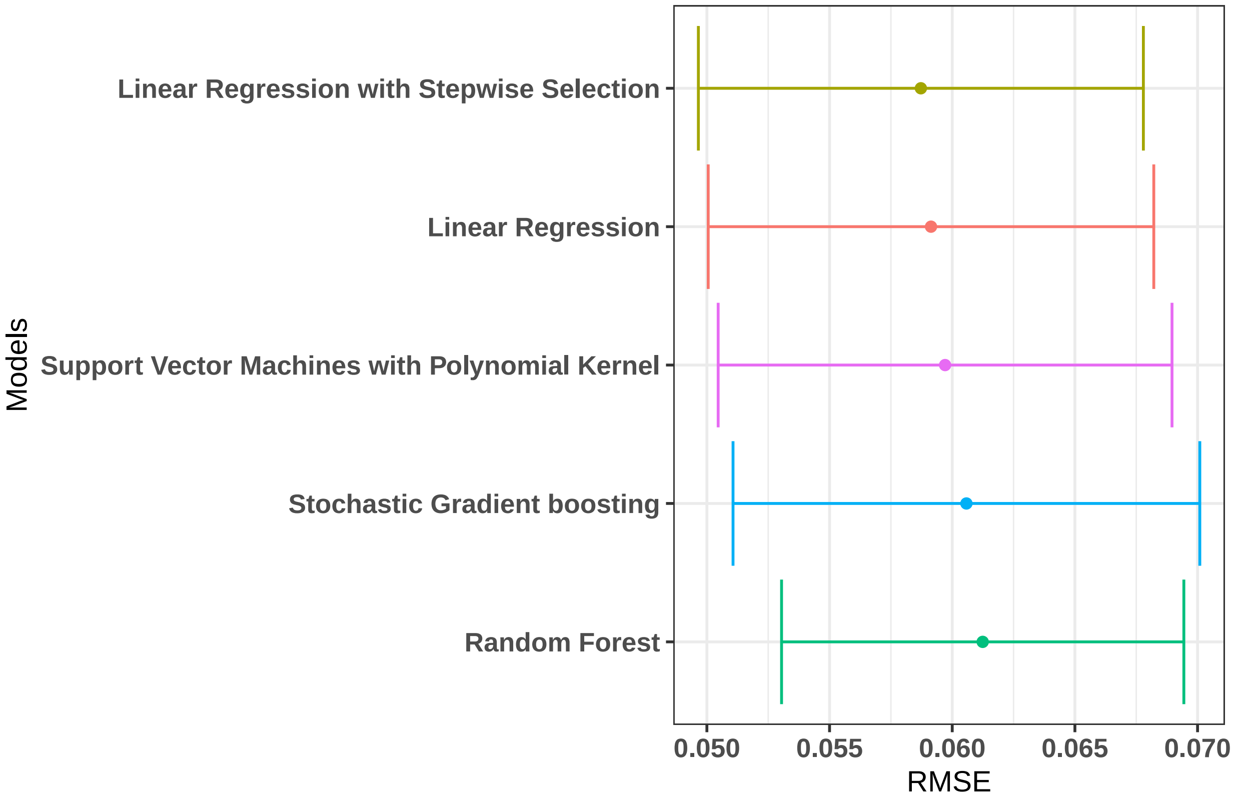
Figure 11c: Model Fit

Figure 11d:Model Residual

Model Section

The performance of the five models were relatively close, when compared using the RMSE. It ranged from a low of 0.0587 for the Linear Regression with Stepwise Selection model to a high of 0.0612 for the Random Forest model (Figure 12).

Figure 12: Best RMSE results from repeated 10 Fold CV for 5 Models Trained



A plot of the predicted values of all models relative to the actual values revealed that no model drastically outperformed the other models at all the points (Figure 13). This suggests that a stacked ensemble model that combines all five models may add to the predictive power.

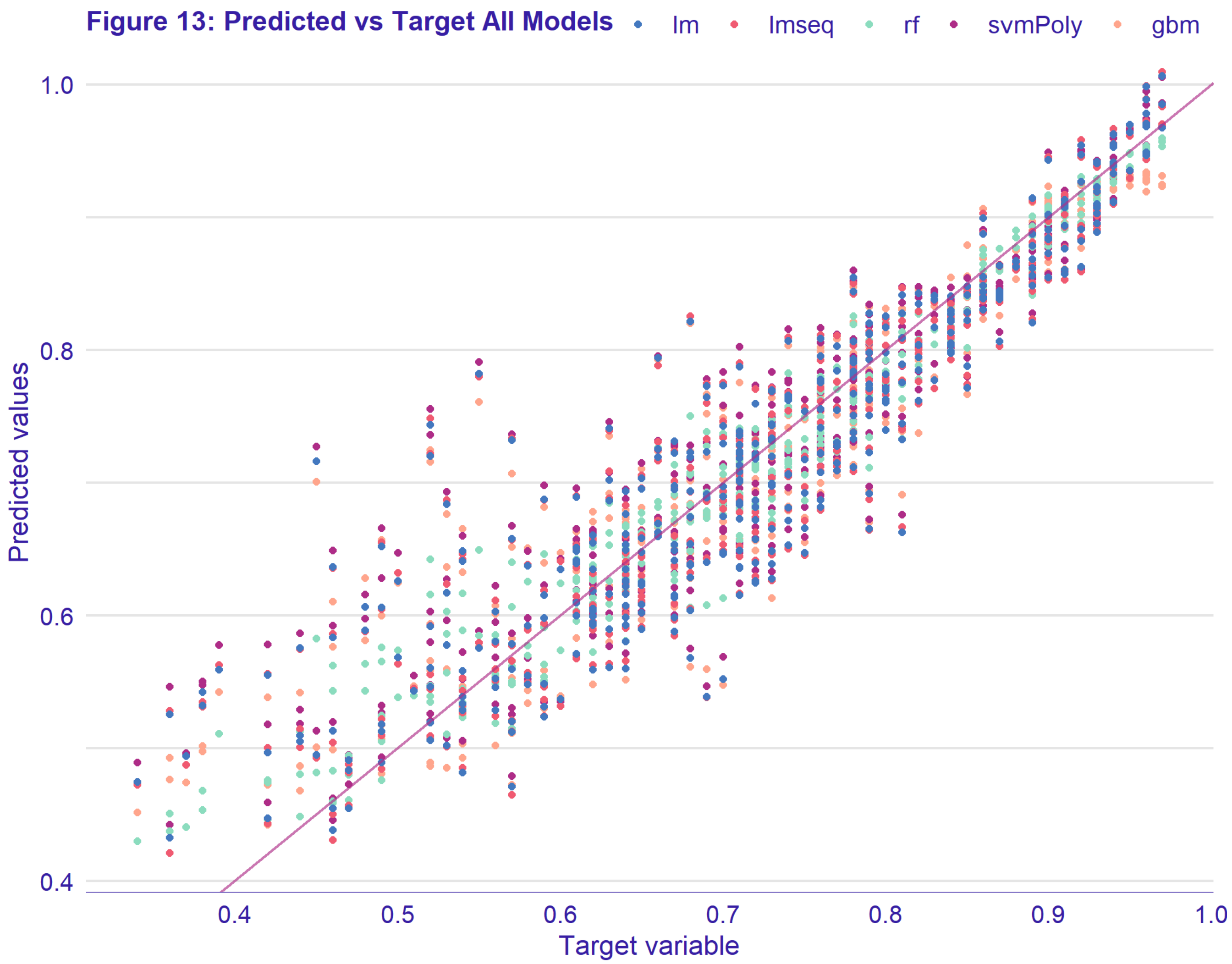


Figure 13: Predicted vs Target All Models

SECTION V: CONCLUSION

Summary

The objective of this paper is to predict a student's perceived chance of admission into a graduate program. The predictors that were used to predict chance of admission were, the student's undergrad GPA, GRE score, TOEFL score, Letter of recommendation, statement of purpose, whether or not the student had research experience and the rating of the student's undergrad university.

A dataset (admission predict) consisting of 500 observations was used in the analysis. The dataset was split into a training dataset and test(hold out) dataset, representing 80% and 20% of the original dataset,

Five machine learning algorithms were trained and tuned on the training dataset using 10 fold cross validation repeated 10 times. The model that produced the lowest RMSE was then used to predict the test dataset.

The RMSE using the 10-fold cross validation repeated 10 times yielded an RMSE that ranged from 0.0587 for the Linear Regression with Stepwise Selection model and 0.0612 for the Random Forest model.

The Linear Regression with Stepwise Selection model was then applied to the test/hold out dataset which produced an RMSE of 0.0639. The relative closeness of the RMSE suggest that the chosen model was not overfitted.

Limitation

One factor that may have contributed to relatively simple models such as linear regression and stepwise linear regression outperforming more complexed machine learning models (e.g. Random Forest) is the relatively small sample size. The larger the dataset the better the performance of more complexed models ceteris paribus.

Further Extension

Further work on this project may involve collecting more data on the variables of interest in addition to the ranking of the graduate program that the student is applying to, which may have a significant influence on their perception of being admitted. For example, applying to Harvard University verus a lower ranked school with the same academic profile may have a different perceived chance of admission.

Additionally, utilize an algorithm that uses v-fold cross validation to combine the different models using a stacked ensemble which has been found to improve predictive performance(van der Laan, Polley, and Hubbard 2007).

REFERENCE

Acharya, Mohan S, Asfia Armaan, and Aneeta S Antony. 2019. "2019 International Conference on Computational Intelligence in Data Science (Iccids)." In. IEEE. https://doi.org/10.1109/iccids.2019.8862140.

Chen, Shyh-Huei, Jielin Sun, Latchezar Dimitrov, Aubrey R. Turner, Tamara S. Adams, Deborah A. Meyers, Bao-Li Chang, et al. 2008. "A Support Vector Machine Approach for Detecting Gene-Gene Interaction." *Genetic Epidemiology* 32 (2): 152–67. https://doi.org/10.1002/gepi.20272.

Elith, J., J. R. Leathwick, and T. Hastie. 2008. "A Working Guide to Boosted Regression Trees." *Journal of Animal Ecology* 77 (4): 802–13. https://doi.org/10.1111/j.1365-2656.2008.01390.x.

García-Magariños, M., I. López-de-Ullibarri, R. Cao, and A. Salas. 2009. "Evaluating the Ability of Tree-Based Methods and Logistic Regression for the Detection of Snp-Snp Interaction." *Annals of Human Genetics* 73 (3): 360–69. https://doi.org/10.1111/j.1469-1809.2009.00511.x.

Greenwell, Brandon,M., and Bradley,C. Boehmke. 2020. "Variable Importance PlotsAn Introduction to the Vip Package." *The R Journal* 12 (1): 343. https://doi.org/10.32614/rj-2020-013.

van der Laan, Mark J., Eric C Polley, and Alan E. Hubbard. 2007. "Super Learner." *Statistical Applications in Genetics and Molecular Biology* 6 (1). https://doi.org/10.2202/1544-6115.1309.