Project Report

On

University Selection Recommendation

By Group No: 4

Name	Enrollment No
Anant Golecha	MB18GID255
Md Asif Imam	MB18GID275
Mradu Payal Sharma	MB18GID291
Prayank Kulshrestha	MB18GID293
Priyank jha	MB18GID294
Suneyna Bansal	MB18GID249
Suraj Kumar	MB18GID272
Vicky Deshpande	MB18GID296

Contents

List Of figures

1. Problem Statement

To prediction whether or not a student will be accepted by the University for admission on the ground of available information regarding the respective students.

2. Prepare the data

Preparation of data involves the following activities-

2.1 Checking for the null values: Many Features has significant Null Values

- State Province (1.2% null values)
- Admission population description (3.4% null values)
- Gender (0.02% null values)
- Class Rank, Class Size, Class Rank Percentile (81% null values)
- ACT Score and SAT Score also have null values
- Institutional Aid offered (41% null values)
- GPA (9% null values)

We observed class Rank, Class Size, Class Rank Percentile have 81% of null values so dropping these feature was the best option available. High School GPA Grades can be imputed by any central tendency strategy. SAT and ACT score, Institutional Aid offer need to be analyse further.

	Academic Period	Unique ID	State Province	Student Population	Application Date	Admissions Population Description	Residency Description	College Description	Major Description	Gender	...	Common Application	College Online Application	Corr Applic Up
column type	int64	int64	object	object	object	object	object	object	object	object	...	int64	int64	
null values (nb)	0	0	491	0	0	1360	0	0	0	9	...	0	0	
null values (%)	0	0	1.2449	0	0	3.44819	0	0	0	0.0228189	...	0	0	

Figure 1 : Null Values and Data Types

2.2 Checking for duplicate values:

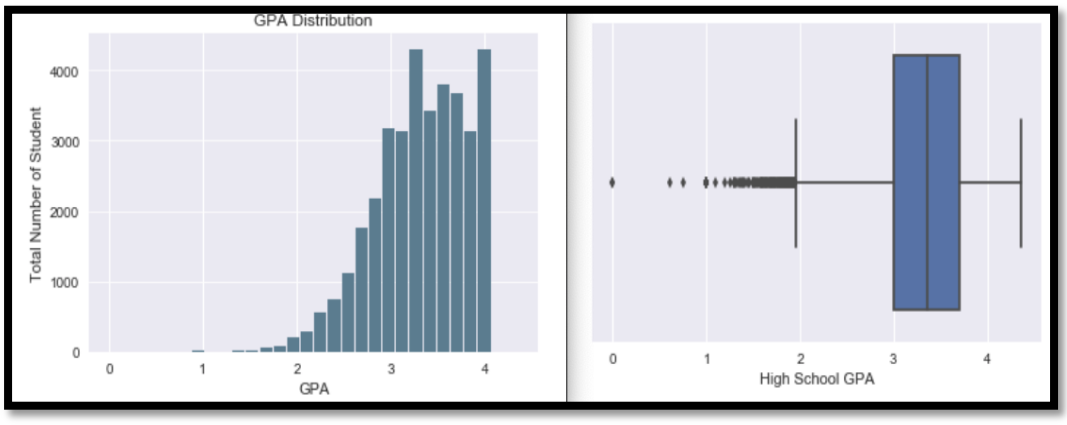
No duplicate values found in our Data.

2.3 Exploring the categorical features

- Pre veterinarian feature has only one category so we can drop this feature because this feature has no significance in model building.
- Admission population description feature has many minor categories such as bridge transfer early decision. So we convert all into a single category.
- Similarly residential description feature also has many minor categories such as international, local community, undeclare. So we converted all these into single category.

2.4 Analysis of High School GPA, Institutional Aid offered, SAT & ACT Score

- GPA score distribution is skewed and not normally distributed and we also observed GPA score has many outliers also. So we removed all the outliers and we observed our data is approximately normally distributed.
- Institutional Aid offer. This feature had 41% null values but we further analyze out of 41% students around 88% students not applied for any kind of financial aid. Such as merit based or need based. So we imputed this feature with zero. And checked the distribution of institutional aid offer and the data was normally distributed and skewed.



Skewed Data **Outliers Removed**

Figure 2: Normally distributed Skewed data after outlier removal

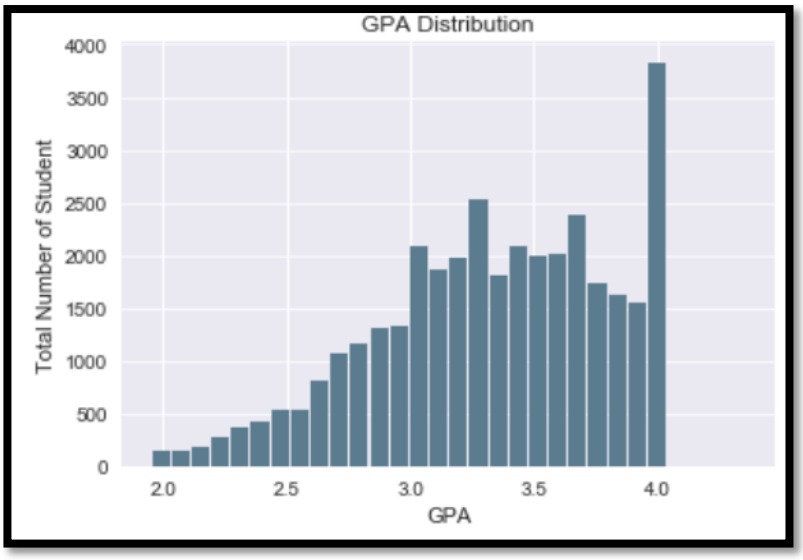


Figure 3: High school GPA Distribution after Outlier Removal

2.5 Date Time Feature Treatment:

We have 2 features in Date Time data type- Academic period and application date.

So we extract year and month information from these features and removed original feature.

2.6 PCA analysis over the SAT or ACT scores

PCA analysis has been done on SAT and ACT score features (total 8 features). And finally we selected three principle components with 97% explained variance.

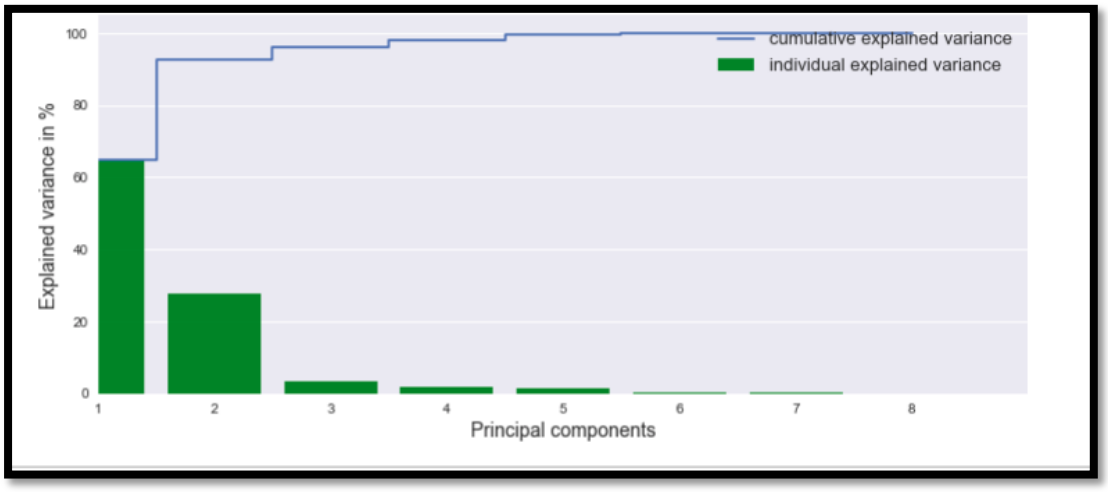


Figure 4 : PCA Component Analysis

2.7 Co-relation Matrix Analysis

- Application year has a high co-relation has other independent feature. So we dropped this feature.

- Merit based financial aid has high co-relation with need based financial aid so we dropped this feature also.

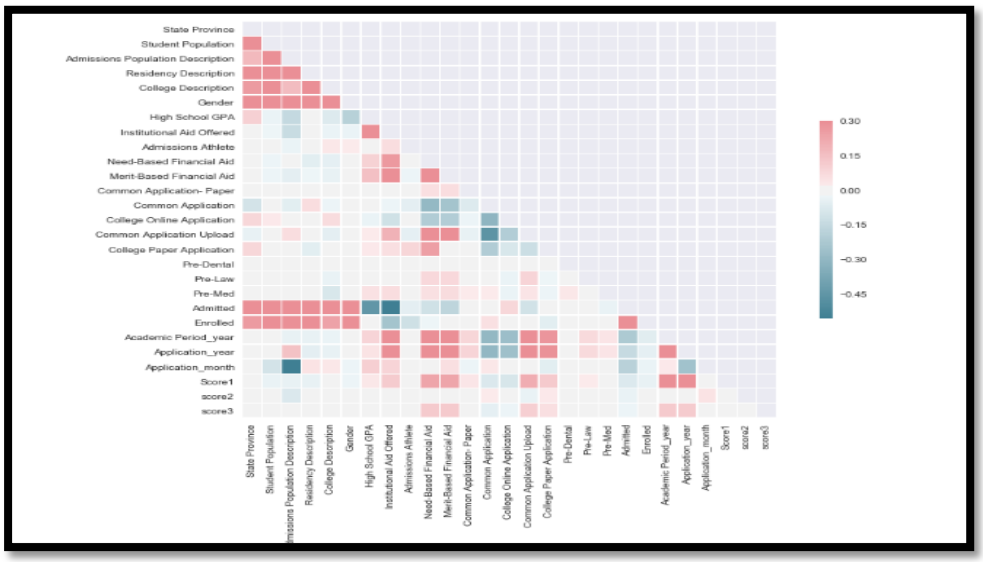


Figure 5 : Correlation Matrix

3. Explore the data

- We can see Pennsylvania has 35.1% of the application followed by New Jersey 13.1%. Under top 15 states with most no of Application

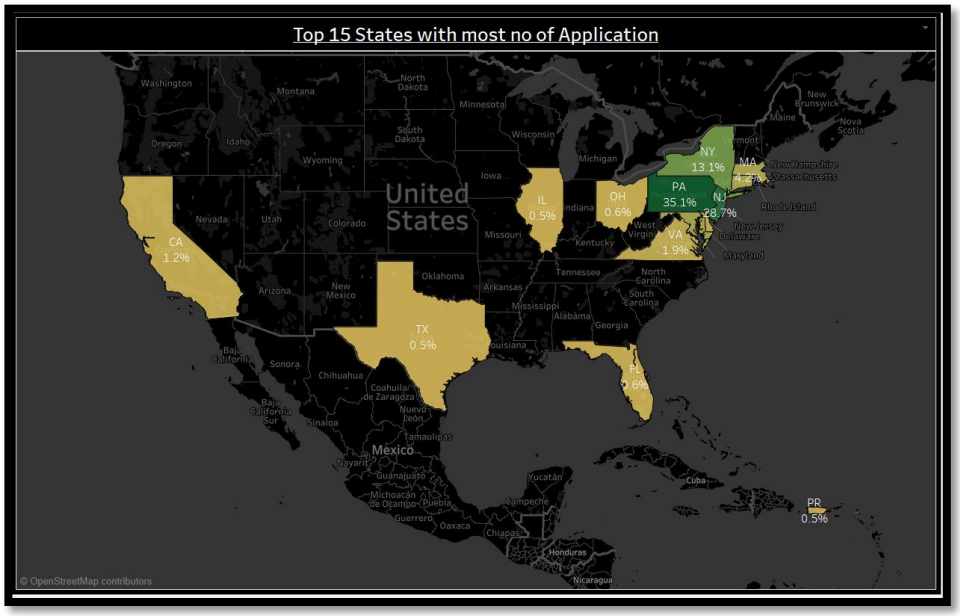


Figure 6 : US state wise candidate acceptance distribution

- In our Data male and female has almost same proportion but female acceptance rate bit higher than male.

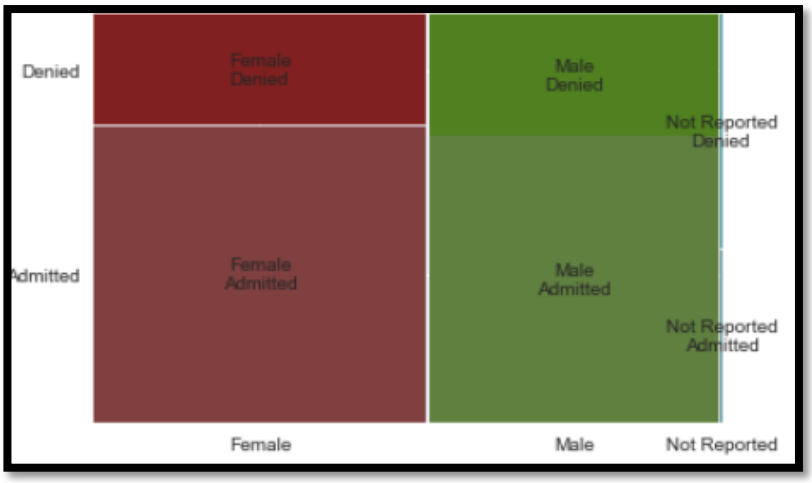


Figure 7: Gender Based Distribution of acceptance Rate

- Student those applied under sport quota or students applied for merit based financial aid have higher chance of acceptance.

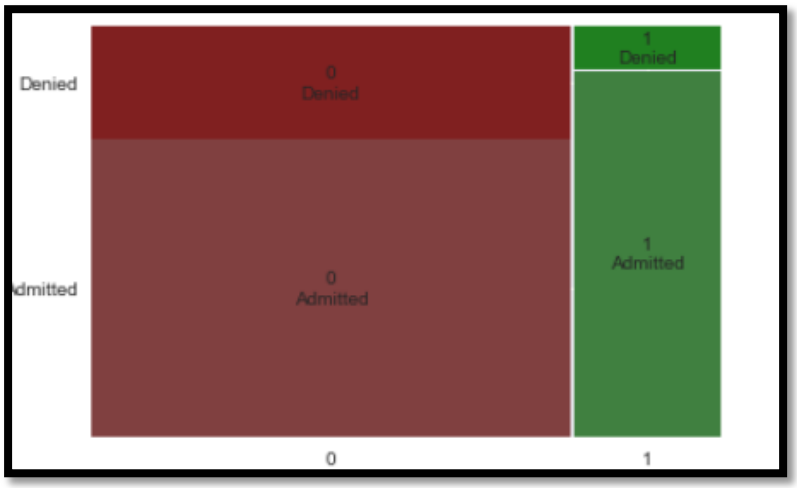


Figure 8: Merit Based Financial Aid Acceptance Distribution

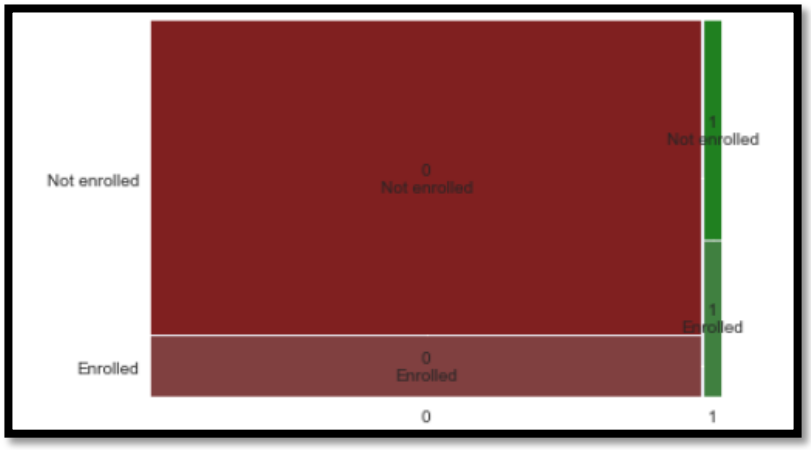
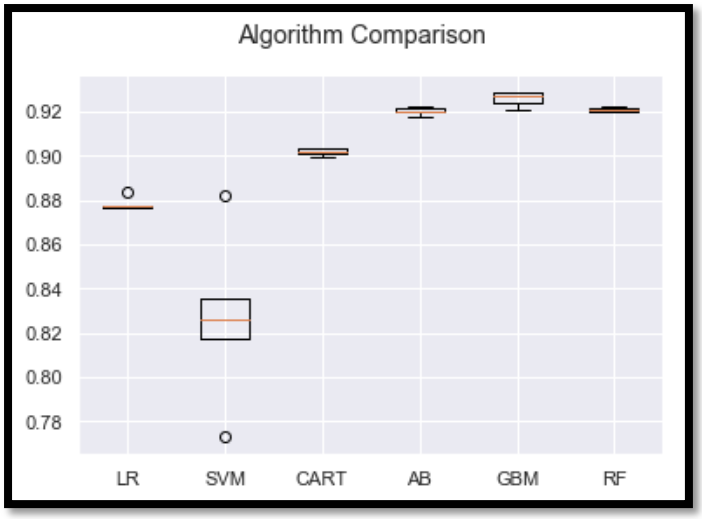


Figure 9: Sport Quota Based acceptance Rate

4. Build models

To check which algorithm will do well, we tried 3 baseline and 3 ensemble techniques. For spot checking, comparison was made with 5 fold cross-validation using accuracy metric (Figure 1). 3 baseline techniques used were Logistic Regression, SVM, and Decision Tree. Three ensemble techniques used were Random Forest, Ado Boost and Gradient Boosting.

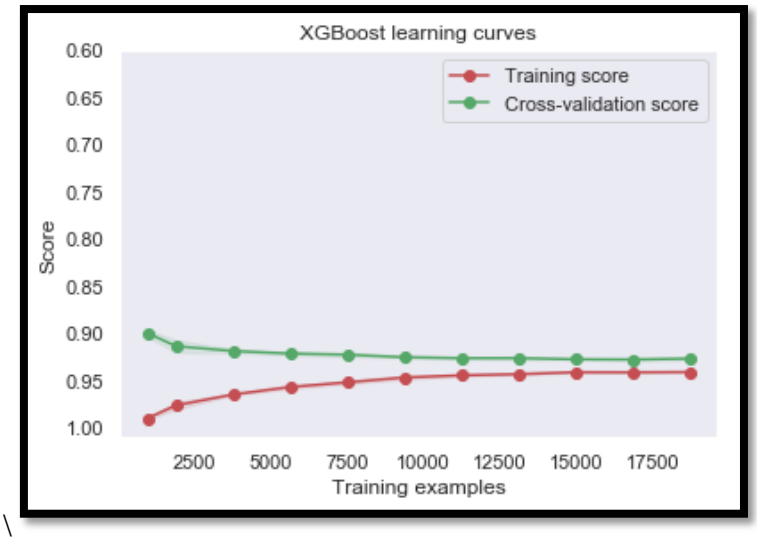


Algorithm	Mean	Standard Deviation
LR	87.84	-0.28
SVM	82.69	-3.51
CART	90.19	-0.15
AB	92.03	-0.17
GBM	92.57	-0.31
RF	92.09	-0.11

Figure 10: Algorithm comparison with 5 fold cross validation using accuracy metric

As evident from the above figure that gradient boosting outperforms as compared to other. However, all ensemble techniques including baseline decision tree seems to have very similar accuracy scores.

In order to further check for any over fitting resulting high accuracy of 93% for gradient boosting, we plot learning curves at train size proportions of 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1 (Figure 2). On this curve, we can see that the train and cross-validation curves converge towards the same limit when the sample size increases. This is typical of modeling with low variance and proves that the model does not suffer from overfitting. Also, we can see that the accuracy of the training curve is correct which is synonymous of a low bias. Hence the model does not underfit the data.



\

Figure 11: Learning curve for gradient boosting

We further used GridSearchCV function to tune the hyperparameters as "learning_rate", "min_samples_leaf", "max_depth" and "n_estimators". Best Parameters found are following

- learning_rate: 0.2
- max_depth: 8
- min_samples_leaf: 9
- n_estimators': 10

However kfold accuracy (94%) was not increased significantly.

Random Forest also performed well with data with Accuracy ~93.25% with K-fold Cross validation and grid search. Grid search function gives best parameter are Max_depth of tree is 10 and Max_features 12.

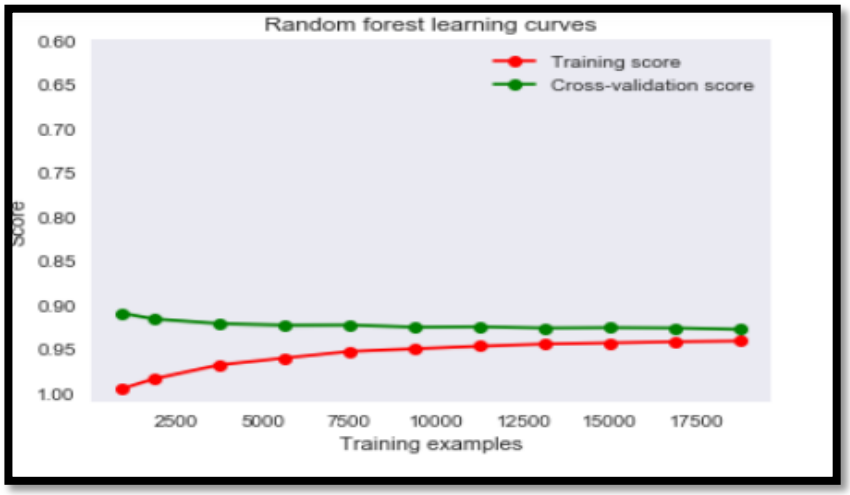


Figure 12: Random Forest Learning Curve

5. Validate Models

Model was validated using k-fold cross validation method as the sample size is not a constraint for us. Thus we choose k-fold =5 and evaluate accuracy for six different algorithm (see Figure 1). Gradient boosting was choosen and further Hyperparameters tuning was done. Again accuracy was checked using 5-fold cross validation. We also made confusion matrix with TN = 7334, TP = 2058, FP = 412 and FN = 241. Please see below mention classification report for precision, recall and F1 score.

precision	recall	f1-score	support	count
0	0.97	0.95	0.96	7746
1	0.83	0.9	0.86	2299
micro avg	0.93	0.93	0.93	10045
macro avg	0.9	0.92	0.91	10045
weighted avg	0.94	0.93	0.94	10045

Figure 13: Classification report for gradient boosting

6. Results, Discussion and Implementation

We got a high accuracy of 94% with gradient boosting algorithm. For further implementing the algorithm, a user interface can be designed for university administration where they can see the list of all applied applicant with a ‘compute score’ button. Pressing the button would call a Rest API which would compute the score for that particular user and provide its class with probability score. Moreover if user wants they can ask for ‘compute for all’ which will give the score profile for all students. Eventually this data can be integrated to training module. Website can also give display for statistics for the results.