Predicting US Graduate Program Admission using User Profiles from Edulix.com

Group 31: Shloak Aggarwal 2017107
Syed Ali Abbas Rizvi 2017114
Tanish Jain 2017115

Data (Source: GitHub + Edulix.com)

	Program	Research Exp	Industry Exp	Intern Exp	Journal Pubs	ConfPubs	TOEFL Score	TOEFL Essay	greV	greQ	greA	Topper CGPA	CGPA	UG label	ranking
0	0	9	24	3.0	0.0	1.0	114.0	0.0	157.0	164.0	4.0	9.50	8.860	734	72
1	0	0	0	0.0	0.0	0.0	104.0	0.0	152.0	162.0	2.5	8.80	7.760	469	167
2	0	0	0	0.0	0.0	0.0	111.0	0.0	153.0	163.0	4.0	9.00	8.417	899	129
3	0	0	0	0.0	0.0	0.0	104.0	0.0	510.0	730.0	4.0	8.50	7.200	469	344
4	0	0	0	0.0	0.0	0.0	0.0	0.0	139.0	161.0	3.0	9.70	8.000	1039	501
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
47514	0	0	0	0.0	0.0	0.0	0.0	0.0	510.0	740.0	3.0	7.60	6.212	662	35
47515	0	0	0	0.0	0.0	0.0	98.0	27.0	152.0	165.0	4.0	9.43	9.110	907	129
47516	0	0	0	0.0	0.0	0.0	95.0	0.0	145.0	164.0	3.0	9.10	7.560	47	484
47517	0	0	0	0.0	0.0	0.0	110.0	0.0	620.0	780.0	3.5	8.30	7.970	228	167
47518	0	0	0	0.0	0.0	0.0	99.0	24.0	151.0	167.0	3.0	9.00	6.500	501	215

47519 rows × 15 columns

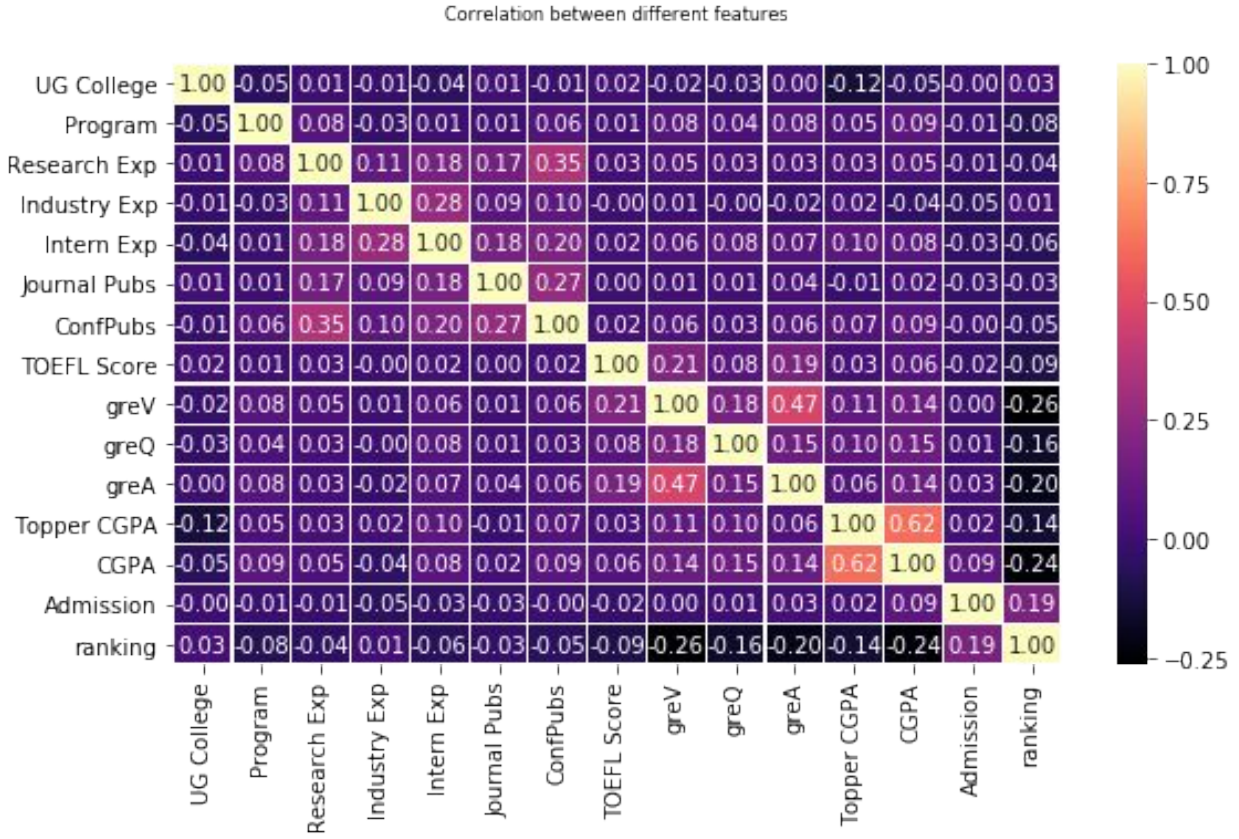
Data cleaning and preprocessing

- Manually cleaned the data.
- Dropping samples which had no value for at least 1 feature left us with almost no samples.
- Using 15 of 26 columns. Discarded columns (after use): 'Username', 'Specialization', 'Major', 'Department', 'User Profile Link', 'Term & Year', 'gmatV', 'gmatQ', 'gmatA', 'CGPA Scale'
- Replaced all string entries with integer entries using LabelEncoder or simple hash functions.
- Replaced the Universities, the students were wanting to take admission in, with their international ranks.
- Used the range values were supposed to be in, to remove outliers. Also used Z-Score to remove some more outliers.

Data cleaning and preprocessing (Cont.)

- Filled all the null values with 0 values, removed samples with 0 as their ‘CGPA Scale’, scaled the CGPA based on the entries in the ‘CGPA Scale’ column and then standardized some of the columns.
- Replaced 0 ‘Topper CGPA’ with 9 and removed entries with CGPA less than 3, ‘Topper CGPA’ less than 5.5 or more than 10 in both cases.
- Replaced 0 entries in greA, greV, greQ columns with the mean of the column.
- Finally shuffled the whole data.
- 53643x26 **---preprocess---->** 47519x15

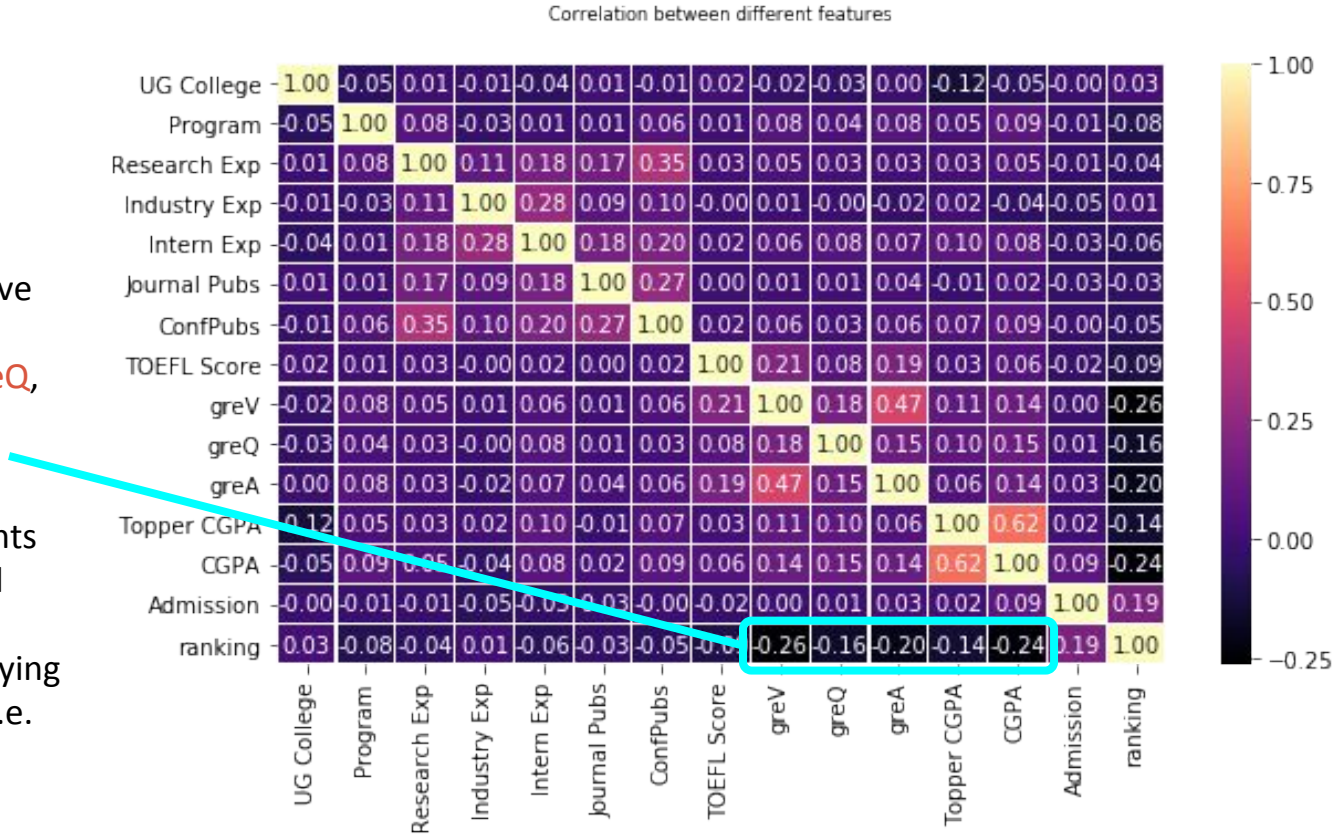
Heatmap


Correlation between different features
UG College 1.00 -0.05 0.01 -0.01 -0.04 0.01 -0.01 0.02 -0.02 -0.03 0.00 -0.12 -0.05 -0.00 0.03
Program -0.05 1.00 0.08 -0.03 0.01 0.01 0.06 0.01 0.08 0.04 0.08 0.05 0.09 -0.01 -0.08
Research Exp 0.01 0.08 1.00 0.11 0.18 0.17 0.35 0.03 0.05 0.03 0.03 0.03 0.05 -0.01 -0.04
Industry Exp -0.01 -0.03 0.11 1.00 0.28 0.09 0.10 -0.00 0.01 -0.00 -0.02 0.02 -0.04 -0.05 0.01
Intern Exp -0.04 0.01 0.18 0.28 1.00 0.18 0.20 0.02 0.06 0.08 0.07 0.10 0.08 -0.03 -0.06
Journal Pubs 0.01 0.01 0.17 0.09 0.18 1.00 0.27 0.00 0.01 0.01 0.04 -0.01 0.02 -0.03 -0.03
ConfPubs -0.01 0.06 0.35 0.10 0.20 0.27 1.00 0.02 0.06 0.03 0.06 0.07 0.09 -0.00 -0.05
TOEFL Score 0.02 0.01 0.03 -0.00 0.02 0.00 0.02 1.00 0.21 0.08 0.19 0.03 0.06 -0.02 -0.09
greV -0.02 0.08 0.05 0.01 0.06 0.01 0.06 0.21 1.00 0.18 0.47 0.11 0.14 0.00 -0.26
greQ -0.03 0.04 0.03 -0.00 0.08 0.01 0.03 0.08 0.18 1.00 0.15 0.10 0.15 0.01 -0.16
greA 0.00 0.08 0.03 -0.02 0.07 0.04 0.06 0.19 0.47 0.15 1.00 0.06 0.14 0.03 -0.20
Topper CGPA -0.12 0.05 0.03 0.02 0.10 -0.01 0.07 0.03 0.11 0.10 0.06 1.00 0.62 0.02 -0.14
CGPA -0.05 0.09 0.05 -0.04 0.08 0.02 0.09 0.06 0.14 0.15 0.14 0.62 1.00 0.09 -0.24
Admission -0.00 -0.01 -0.01 -0.05 -0.03 -0.03 -0.00 -0.02 0.00 0.01 0.03 0.02 0.09 1.00 0.19
ranking 0.03 -0.08 -0.04 0.01 -0.06 -0.03 -0.05 -0.09 -0.26 -0.16 -0.20 -0.14 -0.24 0.19 1.00
UG College
Program
Research Exp
Industry Exp
Intern Exp
Journal Pubs
ConfPubs
TOEFL Score
greV
greQ
greA
Topper CGPA
CGPA
Admission
ranking
1.00
0.75
0.50
0.25
0.00
−0.25

Heatmap

Observation:

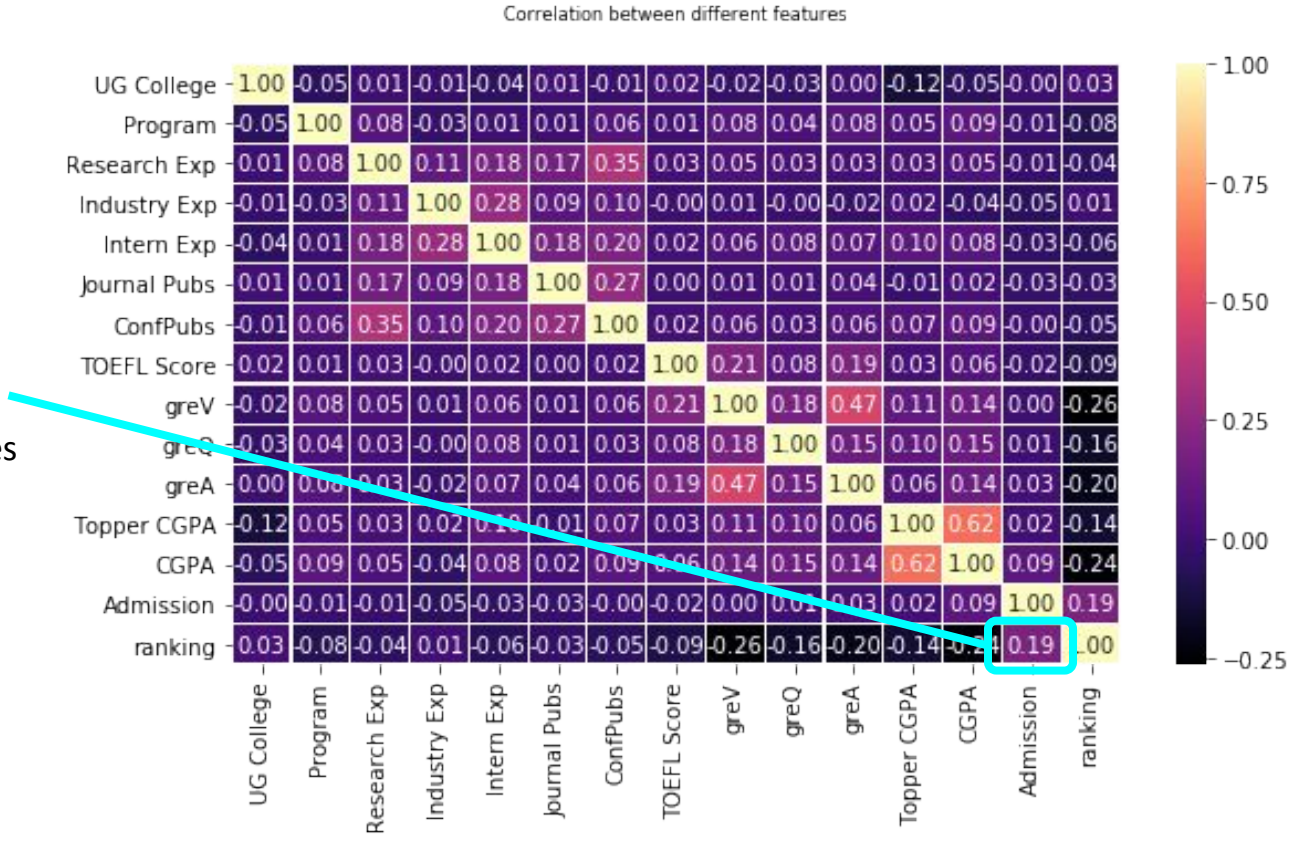
- Relatively high negative correlations between Ranking and greV, greQ, greA, and CGPA.
- Means that the above-average students with higher CGPA and GRE scores are more inclined towards applying to the lower-integer i.e. better ranked universities.



Heatmap

Observation:

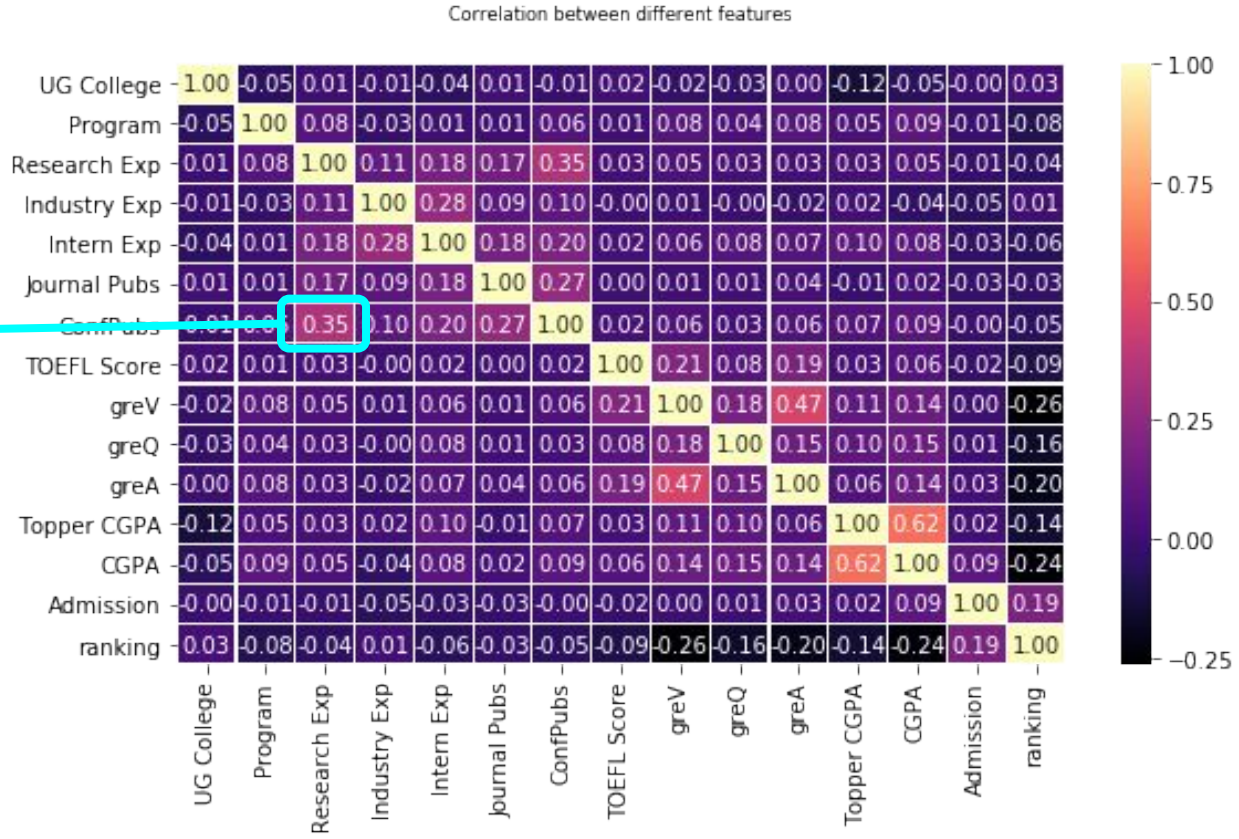
- Relatively high positive correlation between Ranking and Admission
- Lower ranked universities are more lenient about who they accept.



Heatmap

Observation:

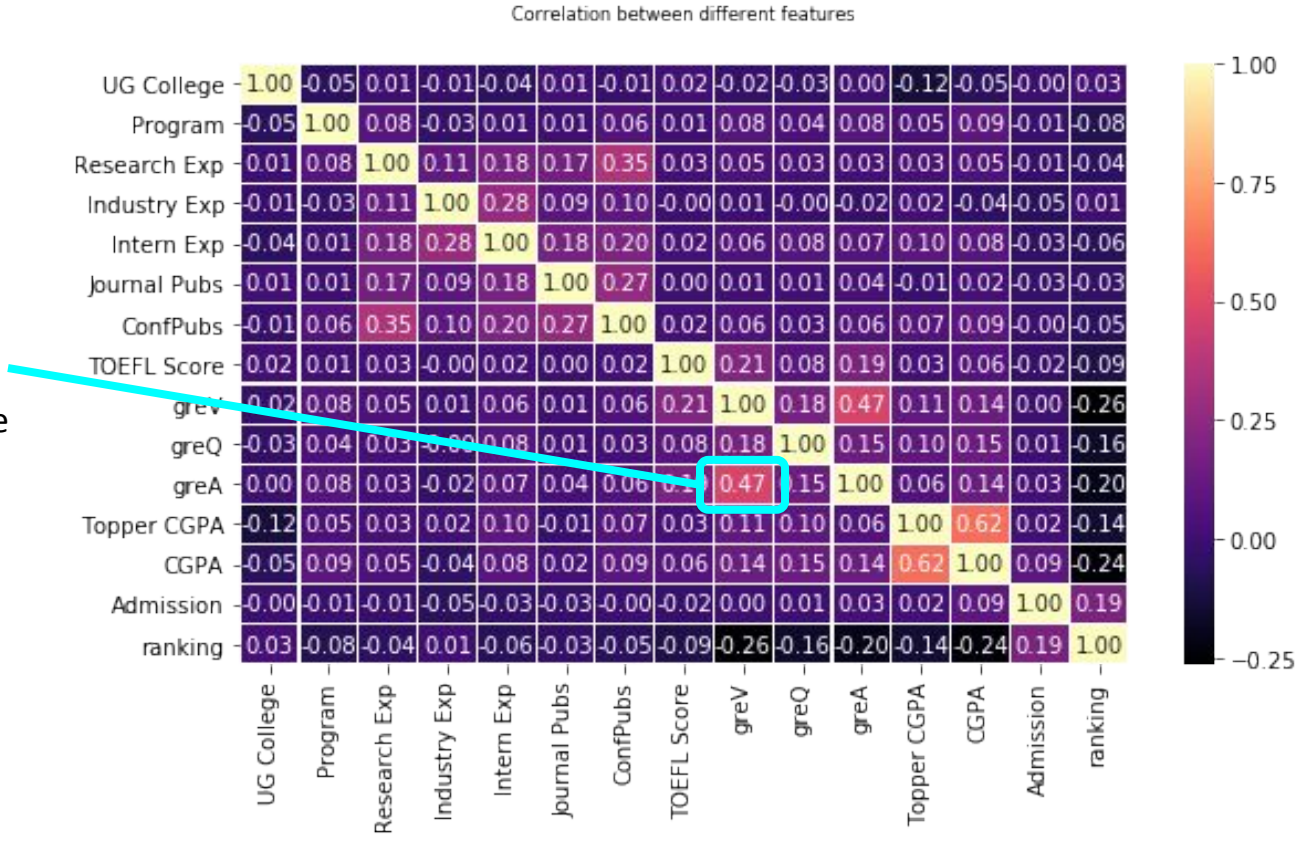
- Conference Publications has a high correlation with Research Experience



Heatmap

Observation:

- greV has a high correlation with greQ, meaning that students who perform well in one section of the GRE are likely to perform well in the other.



Models Used

Logistic Regression with L1 regularisation
Logistic Regression with L2 regularisation

SVM with Poly Kernel, degree = 1
SVM with RBF Kernel
SVM with Linear Kernel

Random Forest Classifier with 10 trees
Random Forest Classifier with 15 trees
Random Forest Classifier with 1000 trees

Multilayer Perceptron (built using Tensorflow)

K Nearest Neighbours (2 neighbours)
K Nearest Neighbours (3 neighbours)

Multi-layer Perceptron (1000 iterations)
Multi-layer Perceptron (2000 iterations)

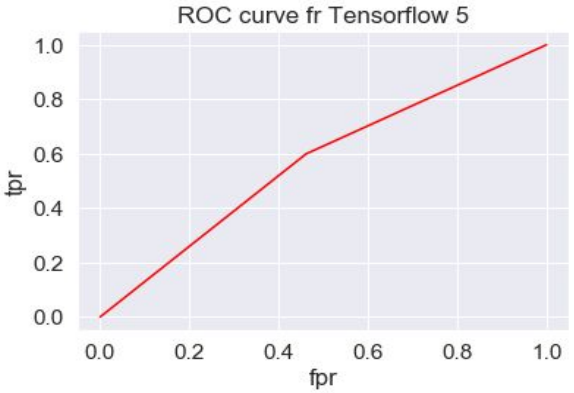
Bernoulli Naive Bayes
Complement Naive Bayes
Gaussian Naive Bayes
Multinomial Naive Bayes

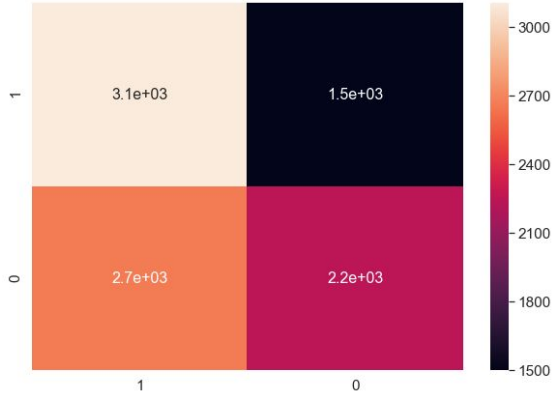
Final Accuracies & F1-Scores:

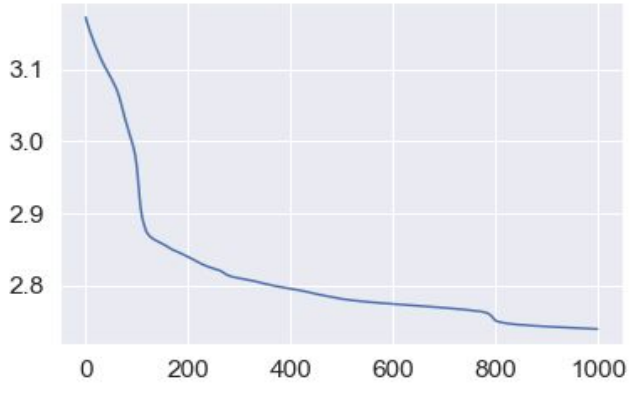
Model	Mean Accuracy	Mean F1-Score
SVM – Linear	0.6064	0.6025
SVM – Polynomial	0.6076	0.5967
SVM – RBF	0.5860	0.6159
Random Forest – 10	0.9117	0.9082
Random Forest – 15	0.9232	0.9243
Random Forest – 1000	0.9267	0.9308
Logistic Regression l1	0.6076	0.6132
Logistic Regression l2	0.5780	0.5899
K-Nearest Neighbours - 2	0.7461	0.6870
K-Nearest Neighbours - 3	0.9232	0.7559
MLP – 1000 iterations	0.5595	0.6047
MLP – 2000 iterations	0.5514	0.3469
Bernoulli Naïve Bayes	0.5290	0.6463
Complement Naïve Bayes	0.5661	0.5503
Gaussian Naïve Bayes	0.5362	0.6557
Multinomial Naïve Bayes	0.5660	0.5504
Tensorflow Neural Network	0.5610	0.5860

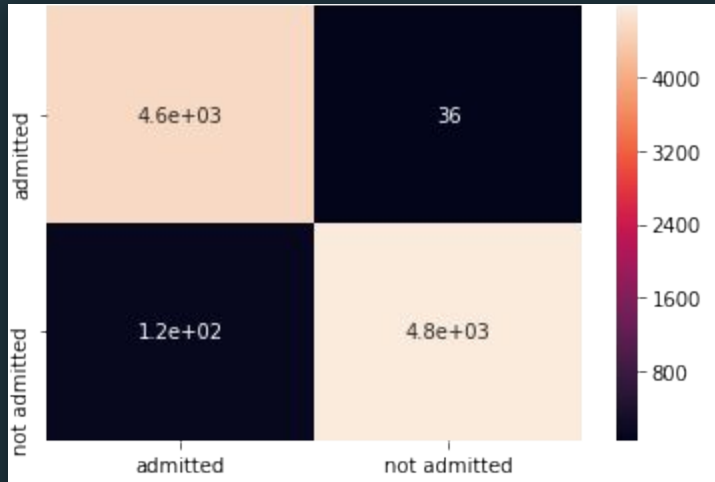
Neural Network created using Tensorflow

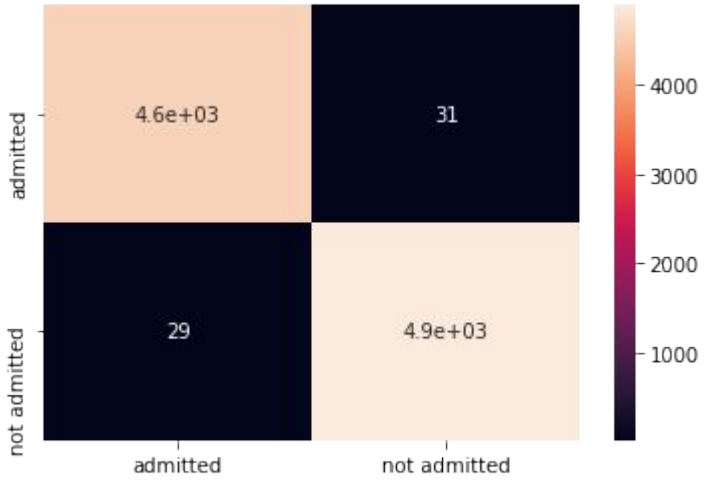
- Custom made Neural Network
- 2 hidden layers with 30 and 45 neurons
- Tanh activation on all hidden layers
- Softmax on Output layer
- Average accuracy = 0.56
- Average f1-score = 0.5860
- ROC Curve, Confusion Matrix and Cost vs Epoch graph respectively:

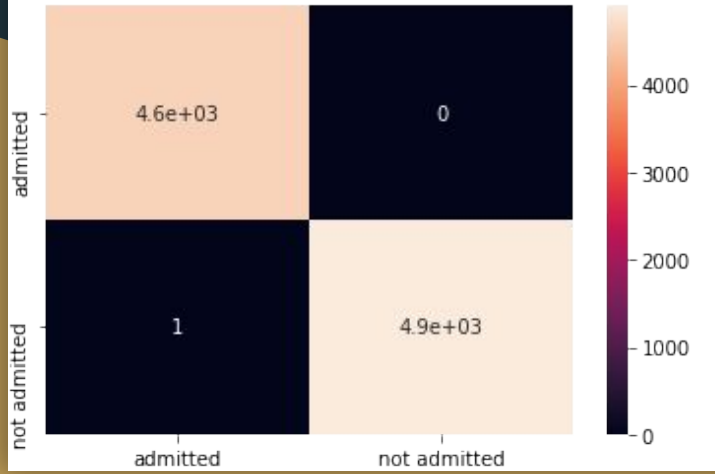






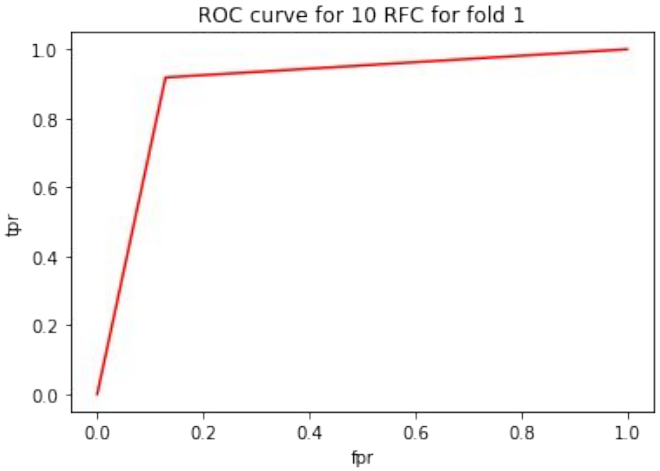


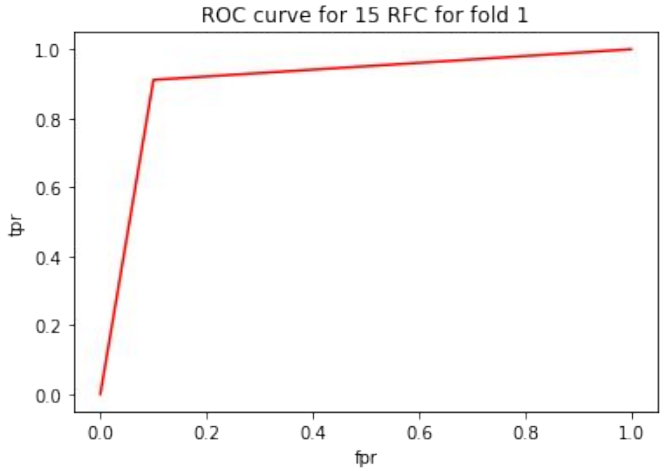


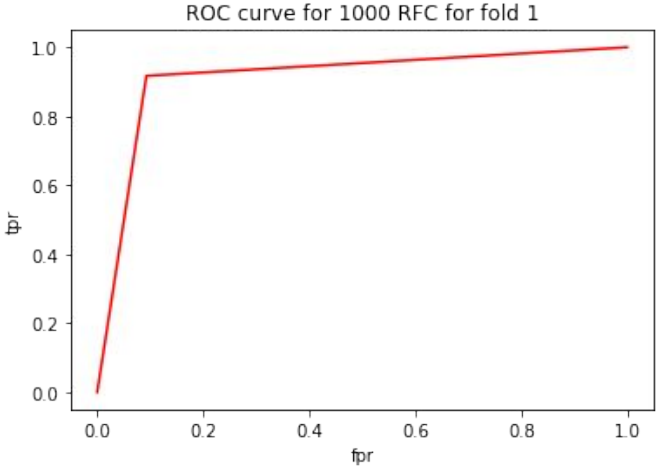


Confusion matrices of Random Forest Classifier with 10, 15 and 1000 trees respectively

Obtained test and train accuracies around 95%

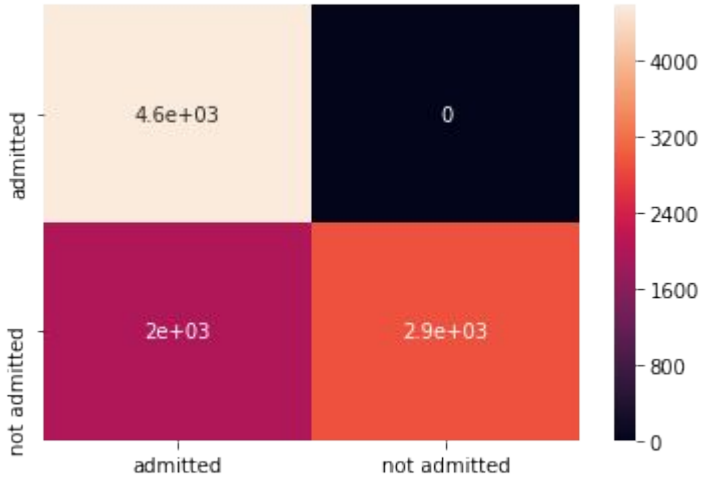


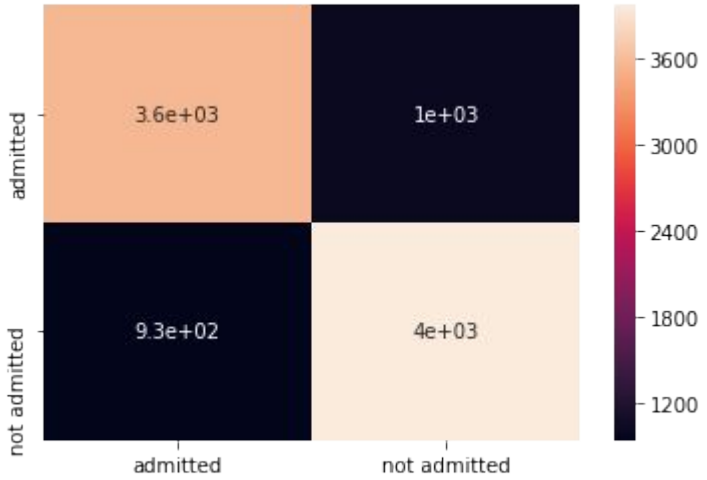




ROC Curves of Random Forest Classifiers with with 10, 15 and 1000 trees respectively

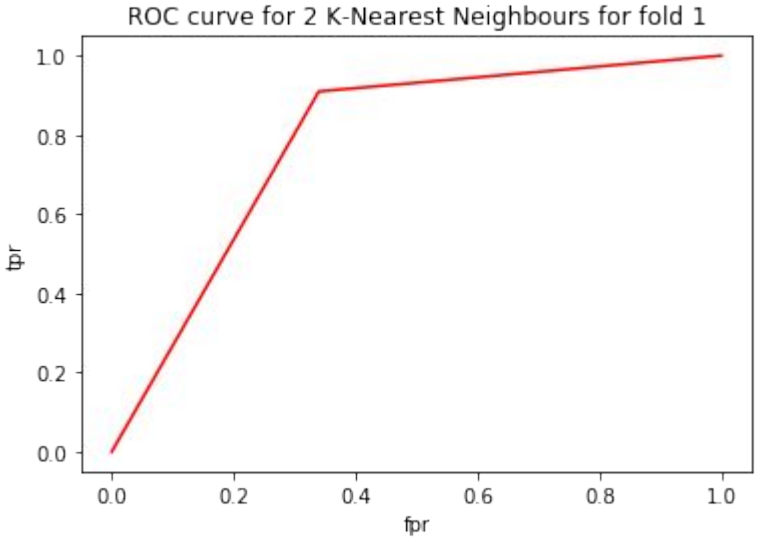
Giving test and train accuracies of around 95%

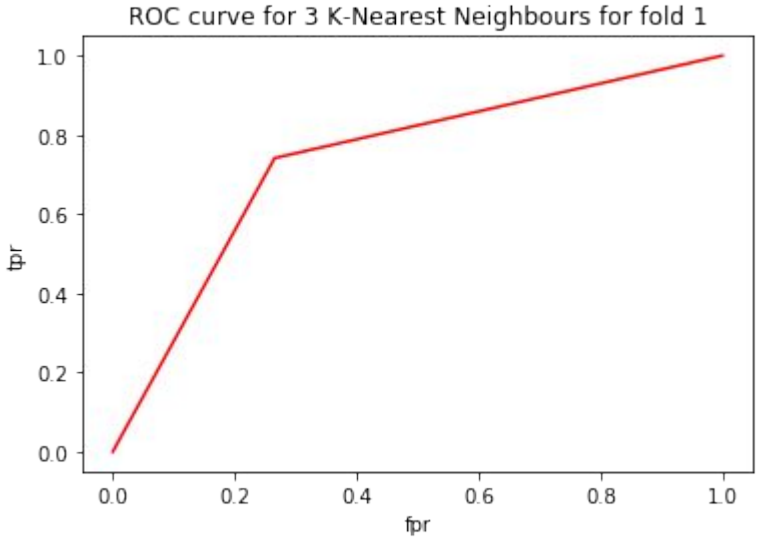




Confusion matrices of K Nearest Neighbour Classifiers with 2 and 3 neighbours respectively

Giving accuracies around 75 %





ROC Curves of K-Nearest Neighbours with 2 and 3 neighbours respectively

Giving testing and training accuracies around 75%

Conclusion

In most models, we obtained accuracies around 60\%, but using K-Nearest Neighbours we obtained accuracies around 75\%, and using Random Forest Classifier we obtained a very high accuracy of around 98\%.

The state-of-the-art accuracy that we came across is around 87\% for similar topics \cite{Waters}. In our study's data, the users who created profiles on Edulix.com did not fully fill all the fields in their profiles. As a result, out data was very sparse. Likewise, other researchers who did similar studies got a different number of samples after they did their own preprocessing.

In most cases, researchers seem to drop a lot of samples as the data is highly sparse. We, on the other hand, only dropped around 6000 samples out of the total 53600. This is why we obtained such high testing and training accuracy using random forest classifier and these accuracies didn't change much even if we changed the number of trees from 10 or 15 to 1000. For the other models, the accuracies remained near 50 to 60\% even after the parameters were varied greatly.