**Report**

**CS Graduate Admission model for Assisting Student:**

We have a labelled dataset (accept & reject) of MS in Computer Science admission profiles scraped from Yocket of 29 Universities in US.

|  |
| --- |
|  |

**Features:** 'gre\_score\_quant', 'gre\_score\_verbal', 'test\_score\_toefl', 'undergraduation\_score', 'work\_ex', 'papers\_published', 'ranking'

**Challenge – Dealing with Imbalanced dataset:**

The dataset seems balanced given only the target label “status”, but we want our model to learn the decision criteria based on the University because it is going to vary for each college. So, we checked for individual counts of target label for grouping our dataset by University and found out imbalance in the dataset. We down-sampled and up-sampled records of some universities using “sklearn.utils.resample” and formed a relatively balanced training dataset of all universities and proceeded with modelling.

**Before resampling:**

|  |
| --- |
| university\_name status  carnegie\_mellon\_university reject 125  accept 37  clemson\_university accept 82  reject 40  george\_mason\_university accept 168  reject 96  georgia\_institiute\_of\_technology accept 87  reject 40  illinois\_institute\_of\_technology accept 268  reject 167  indiana\_university\_bloomington accept 271  reject 72  kansas\_state\_university accept 44  reject 32  michigan\_technological\_university accept 101  reject 37  new\_york\_university reject 179  accept 139  north\_carolina\_state\_university\_raleigh reject 307  accept 281  northeastern\_university reject 1079  accept 574  rochester\_institute\_of\_technology accept 203  reject 138  rutgers\_university\_new\_brunswick reject 196  accept 108  state\_university\_of\_new\_york\_at\_stony\_brook reject 321  accept 281  syracuse\_university reject 280  accept 238  texas\_a\_m\_university\_college\_station reject 221  accept 166  university\_of\_california\_irvine reject 286  accept 114  university\_of\_cincinnati accept 200  reject 42  university\_of\_colorado\_boulder reject 195  accept 146  university\_of\_connecticut accept 38  reject 26  university\_of\_florida accept 60  reject 35  university\_of\_iowa accept 43  reject 16  university\_of\_maryland\_college\_park reject 232  accept 32  university\_of\_north\_carolina\_at\_charlotte reject 195  accept 185  university\_of\_southern\_california accept 28  reject 10  university\_of\_texas\_arlington accept 178  reject 134  university\_of\_texas\_austin reject 220  accept 13  university\_of\_texas\_dallas reject 245  accept 229  worcester\_polytechnic\_institute reject 37  accept 33 |

**After resampling:**

|  |
| --- |
| university\_name status  carnegie\_mellon\_university accept 125  reject 125  clemson\_university accept 125  reject 125  george\_mason\_university accept 168  reject 125  georgia\_institiute\_of\_technology accept 125  reject 125  illinois\_institute\_of\_technology accept 268  reject 167  indiana\_university\_bloomington accept 271  reject 125  kansas\_state\_university accept 125  reject 125  michigan\_technological\_university accept 125  reject 125  new\_york\_university reject 179  accept 139  north\_carolina\_state\_university\_raleigh reject 307  accept 281  northeastern\_university accept 300  reject 300  rochester\_institute\_of\_technology accept 203  reject 138  rutgers\_university\_new\_brunswick reject 196  accept 108  state\_university\_of\_new\_york\_at\_stony\_brook accept 300  reject 300  syracuse\_university reject 280  accept 238  texas\_a\_m\_university\_college\_station reject 221  accept 166  university\_of\_california\_irvine reject 286  accept 114  university\_of\_cincinnati accept 200  reject 125  university\_of\_colorado\_boulder reject 195  accept 146  university\_of\_connecticut accept 125  reject 125  university\_of\_florida accept 125  reject 125  university\_of\_iowa accept 125  reject 125  university\_of\_maryland\_college\_park reject 232  accept 125  university\_of\_north\_carolina\_at\_charlotte reject 195  accept 185  university\_of\_southern\_california accept 125  reject 125  university\_of\_texas\_arlington accept 178  reject 134  university\_of\_texas\_austin reject 220  accept 125  university\_of\_texas\_dallas reject 245  accept 229  worcester\_polytechnic\_institute accept 125  reject 125 |

**Models:**

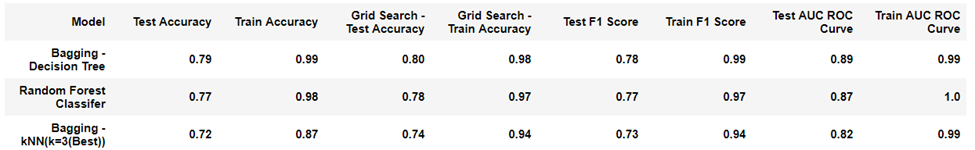
Test-Train split: 0.25

We tried Support Vector Machines, Decision Tree Classifier, Random Forest Classifier, Bagging Model with Decision Tree classifier and kNN, MLP Classifier, GNB Classifier on our data and found out ensemble techniques like Random Forest and Bagging models are giving best results giving upto 80% test accuracy and which we pickled and deployed in our website.

**All model results:**



**Summarizing Best Results:**



* Bagging model with Decision Tree is giving us best after GridSearchCV(CV=5) results with ~80% test accuracy and 0.78 f1 score
* Random Forest Classifier is giving us 2nd best results best after GridSearchCV(CV=5) with ~77% test accuracy and 0.77 f1 score
* Bagging model with kNN(k=3) is giving us 3rd best results best after GridSearchCV(CV=5) with ~72% test accuracy and 0.73 f1 score

**kNN:**

An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors

**SVM:**

SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

**Random Forest Classifier:**

The “forest“ it builds, is an ensemble of Decision Trees, most of the time trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction

**Decision Tree:**

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.

**MLP (Neural Net):**

A multilayer perceptron (MLP) is a class of feedforward artificial neural network. A MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training.

**Boosting:**

Boosting is a machine learning ensemble meta-algorithm for primarily reducing bias, and also variance in supervised learning, and a family of machine learning algorithms that convert weak learners to strong ones. The main variation between many boosting algorithms is their method of weighting training data points and hypotheses. AdaBoost is very popular and the most significant historically as it was the first algorithm that could adapt to the weak learners. It is often the basis of introductory coverage of boosting in university machine learning courses. There are many more recent algorithms such as LPBoost, TotalBoost, BrownBoost, xgboost, MadaBoost, LogitBoost, and others. Many boosting algorithms fit into the AnyBoost framework, which shows that boosting performs gradient descent in a function space using a convex cost function.

**Bagging:**

Bootstrap aggregating, also called bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting. Although it is usually applied to decision tree methods, it can be used with any type of method. Bagging is a special case of the model averaging approach.

**Citation:**

Wikepedia for Machine learning model definitions