Overview of problem

This project is aimed at analyzing the effectiveness and performance of different machine learning algorithms in its ability to classify twitter accounts as bots or humans. For obvious reasons, this is treated as classification problem. The project evaluates the generalization performance and the predictive performance of all the models on future data. Based on the results we find the best suited machine learning algorithm for the given hypothesis space.

Overview of Dataset

- The user account information was obtained from kaggle.
- There are 19 features present in this dataset.
- Additionally, we collected 100 tweets for each user and derived several new features.
- In total, including derived data we had 49 features.

```
id                       575 non-null float64
id_str                   575 non-null object
screen_name              575 non-null object
location                 387 non-null object
description              538 non-null object
url                      383 non-null object
followers_count          575 non-null int64
friends_count            575 non-null int64
listedcount              575 non-null int64
created_at               575 non-null object
favourites_count         575 non-null int64
verified                 575 non-null object
statuses_count           575 non-null int64
lang                     575 non-null object
status                   572 non-null object
default_profile          575 non-null bool
default_profile_image    575 non-null bool
has_extended_profile     575 non-null bool
name                     574 non-null object
bot                      0 non-null float64
In_reply                 568 non-null float64
retweet_count            568 non-null float64
fav_count                568 non-null float64
total_usrmention         568 non-null float64
texts                    568 non-null object
created_at_list          568 non-null object
days_std                 568 non-null float64
hours_std                568 non-null float64
```

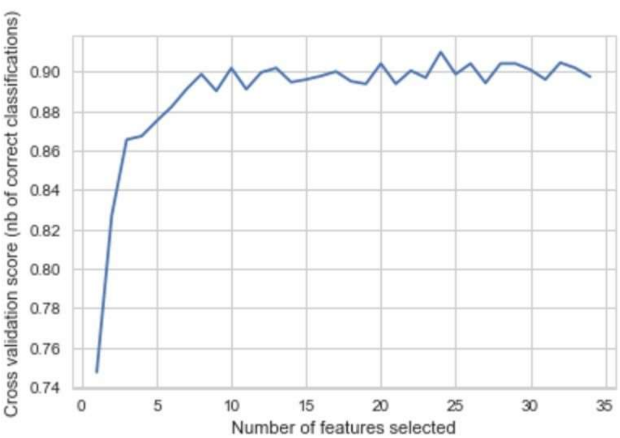
Preview of data set

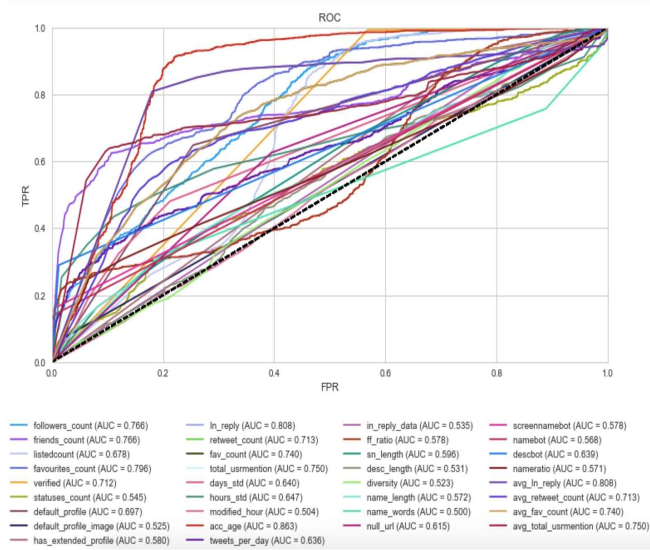
Preprocessing Data

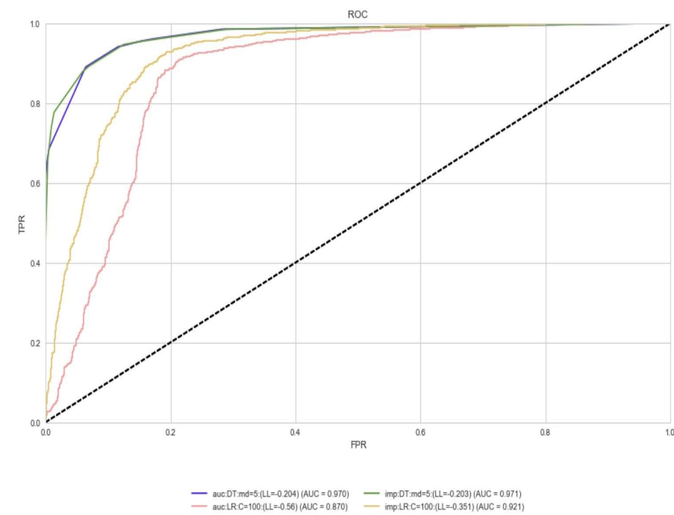
- Missing data was replaced with mean or median of its respective feature
- Categorical data was encoded with discrete numerical values
- Smoothing was done to numerical data
- Extracted time based information and derived several new features
- Data was split into test and train
- Features were standardized.

Feature Selection

- Recursive feature elimination with cross validation was used to select features
- AUC curve for each feature was plotted and feature importance rank was obtained
- Logistic regression and decision trees were used to predict with feature sets from both rfecv and auc ranks.
- The feature set that performed the best was selected.



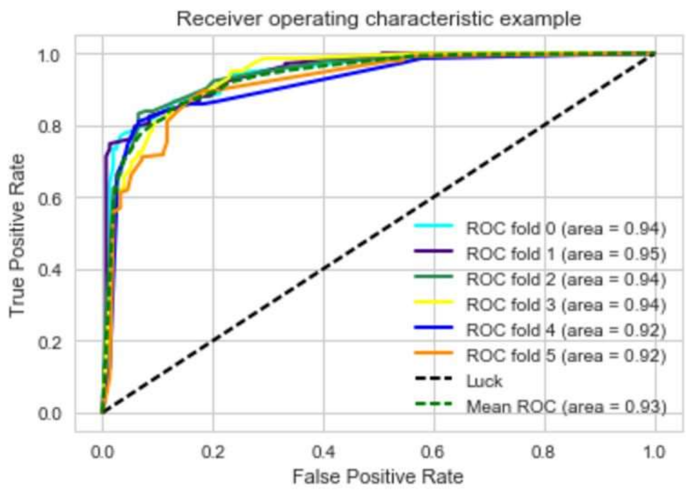




ML Models

Decision Trees

- This was used as a base estimator for our prediction.
- Parameters like max depth of the tree, min impurity split was chosen using grid search.
- It had an test data accuracy of 0.88 and a cross validation mean of around 0.87
- From the confusion matrix it was evident that it was considering more humans to be bots than bots to be humans.It had a false negative of 21 and a false positive of 40
- Despite experimenting with several hyperparameters, the accuracy did not improve further.
- This was an indication that an ensemble technique was required to improve accuracy
- The roc curve and auc for each fold is shown in the below diagram



Logistic Regression

- As this classifier assigns weights to features, we needed to scale the data in order to reduce the influence of high scale features skewing prediction results
- On using gradient descent optimization we obtained a prediction accuracy 0.85 on test data and .79 on cross validation.
- Performing l2 regularization improved the cross validation accuracy to 0.85
- Limited-memory BFGS optimization method was used to find optimal weights with l2 regularization and the accuracy improved to 0.86
- In order to further improve the accuracy, polynomial features were derived with a polynomial degree of 3 and this increased the prediction accuracy to 0.89

Ensemble Techniques

Random Forest

- Random forests with 10 decision trees and and each with a max depth of 3 resulted in an accuracy of .89
- Increasing the max depth of the decision trees to 6 and the number of estimators to 50 resulted in a higher cross validation accuracy of .91

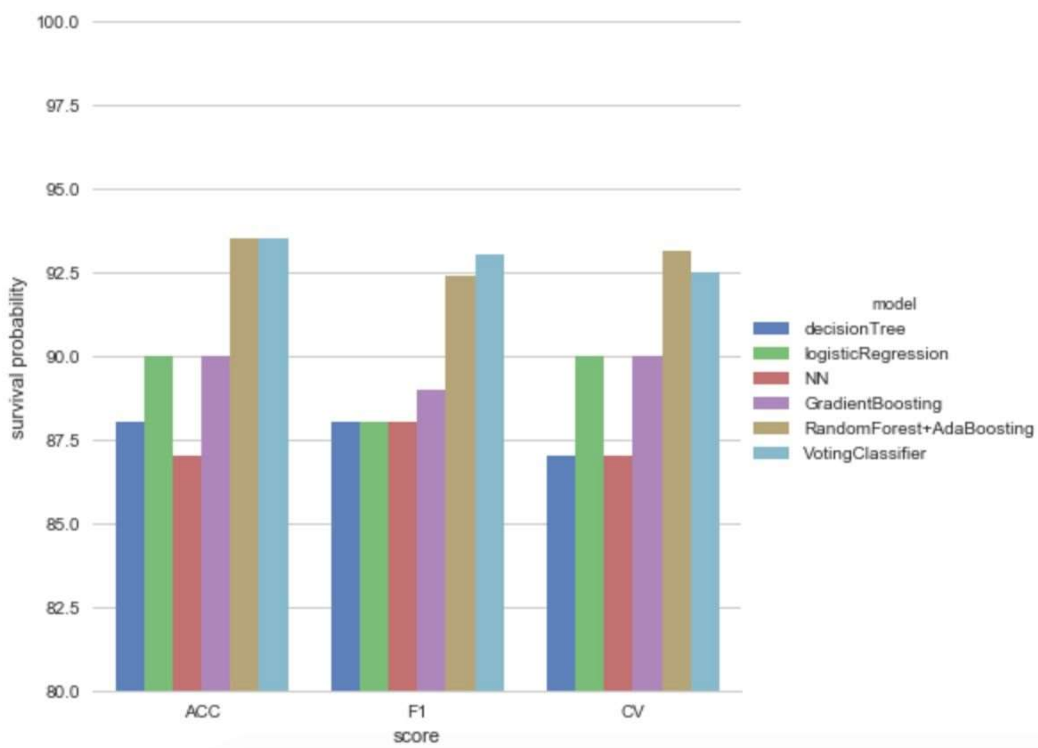
Random Forest with ADA Boosting

- In order to make our model sensitive to noisy data and outliers, adaptive boosting technique was used.
- It resulted in an overall cross validation accuracy of .92

Voting Classifier

- Armed with an ensemble of random forests with ada boosting and logistic regression with gradient boosting, this classifier selects the majority class from the predictions made by all the base classifiers

Conclusion


100.0
97.5
95.0
92.5
90.0
87.5
85.0
82.5
80.0
survival probability
ACC
F1
CV
score
model
decisionTree
logisticRegression
NN
GradientBoosting
RandomForest+AdaBoosting
VotingClassifier