ATP Project

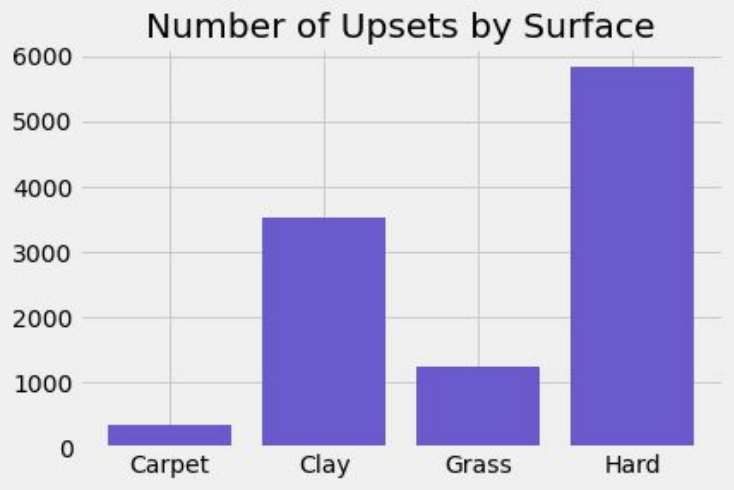
Marcel Colvin

Intro

- Can we create a model that can predict upsets in ATP tennis matches
- A bettor may find our model useful to try and beat odds set by a bookie
- A tennis fanatic may also find our model useful in discovering what features are relevant in an upset.

Methodology

- Data: The data set is about 32,000 tennis matches, played between 2000 and 2018. This data includes features such as surface, tournament round, and the ranks of the players.
- Tools
  - Pandas and Numpy for Feature Engineering and Upsampling
  - Sci-Kit Learn for modeling and normalization
  - Matplotlib for plotting



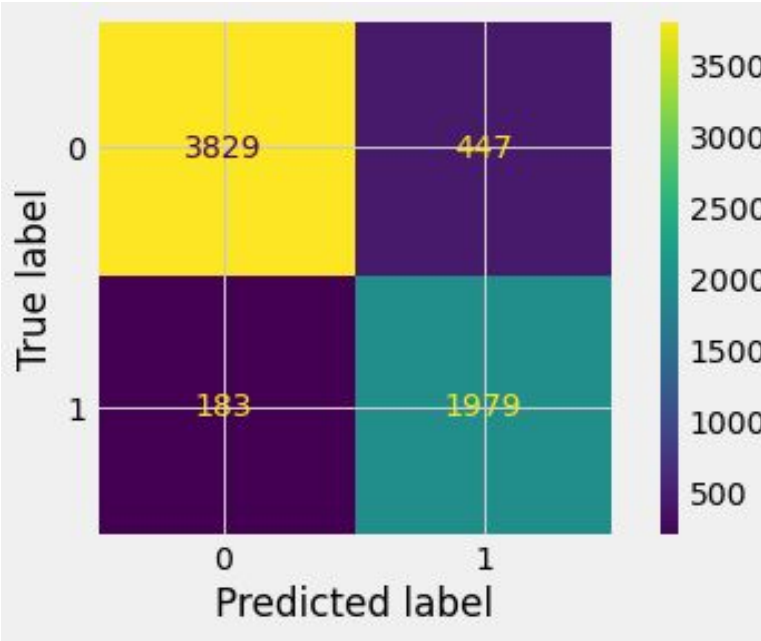
Methodology

- Metrics:
  - Precision is the most important metric for this project, because we want to minimize false positives, but it can be manipulated by models predicting everything as a non-upset i.e. 0/0
  - Used Receiver Operating Characteristic Area Under Curve (ROC AUC) to optimize models
- Models:
  - KNN
  - Logistic Regression
  - Extra Trees
  - Random Forest

Results (initial models)

- KNN performed okay, but poorly compared to other models, because it gives equal weight to all features.
- Extra trees performed poorly on the initial data set before upsampling, by predicting everything as a non-upset
- Logistic Regression was able to perform quite well, but not as well as an optimized tree model

Logistic Regression Confusion Matrix

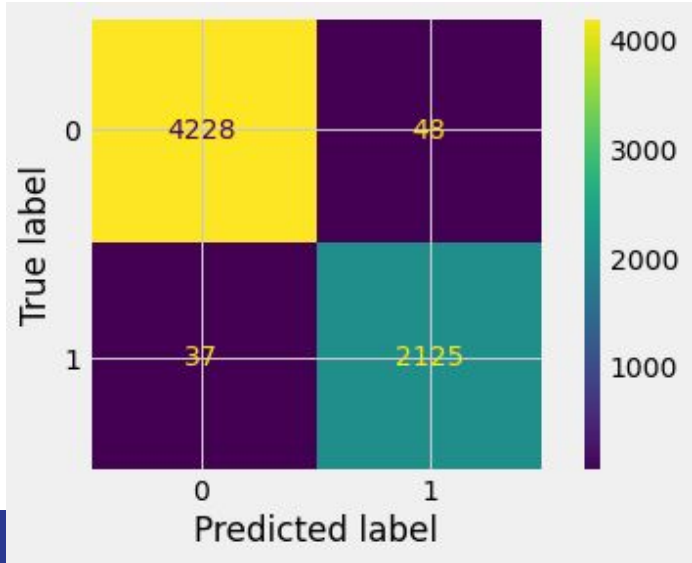


Results (Random Forest Classifier)

- Random Forest performed the best on the initial model
- Performed extremely well after optimizing the hyperparameters with a Grid Search Cross Validation
- Most relevant features were the ranks and elo ratings of the players, the other features seemed to not have an impact.

```
Accuracy: 0.9867971419695558
Precision: 0.9779107225034515
Recall: 0.982886216466235
F1: 0.9803921568627452
```

RFC Confusion Matrix



Conclusions

- The Random Forest model is a relatively good model for predicting tennis upsets in ATP tennis matches.
- The use of bagging and feature randomness in creating the forest of decision trees makes it robust to handle all the features that are involved in this data set

Future Work

- Start working with bookie odds data and trying to create a model that can beat them.
- Work with other sports data to see if we can predict upsets in those cases as well.

Thank you