SHE PLAYS

Submitted By,

Sanjana Rajagopala

1 ABSTRACT

Sports Analytics has reached new levels and the generated data can now be utilized to create winning game strategies. This project involves the creation of a predictive model that will estimate the outcome of the NCAA Women's Basketball tournament based on historical data. It takes support from various techniques of Machine learning, namely Supervised Learning, Random Forest, Logistic Regression, Cross Validation, Regularization etc. to finally meet the goal of predicting the win, lose and draw probabilities of each matchup in the NCAA Women's Basketball tournament.

2 INTRODUCTION

'She Plays' was conceived after developing an inclination towards Sports Analytics due to continuous discussions between the project team and the eventual realization of its deficiency in the field of Women sports games. After careful research and exchange of ideas with the Professor we came across the – 'Google Cloud & NCAA ML Women's Basketball Competition 2018' dataset on Kaggle which involved predicting the outcomes of March Madness during the 2018 NCAA Women's Basketball Championships.

There were three main milestones in this project.

First, **Data Acquisition, Munging and Exploration** which involved collecting data records from the source Google Cloud NCAA ML competition. As the required attributes were spread across different files, integrating them into a consolidated dataset was an imperative task. This project involved identifying fields over which joining could be done efficiently. The Data exploration phase involved the derivation of useful and interesting relationship between the attributes, and thus, understanding the consolidated data.

Second, **build a Predictive model based on historical data.** This task involved the application of Machine learning concepts like Supervised Learning, Logistic Regression and Random Forest to build a predictive model. We built a variety of models using different combination of features. With the help of Regularization, Cross Validation and Feature Engineering, overfitting was avoided, and the model performance was improved.

Third, **use the developed Predictive model for Prediction of 2018 championship results.** The above created model is used to predict the probability that the team with lower id beats the team with higher id in each match. For example, if the tournament has 64 teams, then the model will predict (64*63)/2=2016 match ups. Here, if the team1 id is 3102 and team2 id is 4820, then the probability of team1 beating team2 in the match is predicted for the season 2018.

3 MATERIALS AND METHODS

3.1 Data Import and Preprocessing

The data for this project was obtained from Kaggle website. It encompassed 20-year historical data of women's basketball tournament spread across multiple CSV files. These were imported into data bricks for preprocessing. This involved the integration of information such as mapping between team ID's and team names, seed information of all teams, slot information of each season. Thus, a merged data set of historical data for the years 1998-2017 was created.

Using the additional information regarding individual match characteristics for the matches that happened between the years 2010-2017 another merged data set was created.

Finally, a third data set containing match ups for the year 2018 was created which will be used for final prediction.

SPARK SQL DATAFRAMES, NUMPY and PANDAS were used for data preprocessing and cleaning.

We created derived variables like win percentages for different year ranges (1998-2005,2006-2010,2011-2015,2016-2017), various ratios of match characteristics and included in the data frame accordingly.

3.2 Machine learning Algorithms and Models

As the project deals with a classification problem i.e. predicting winners of each match, the following algorithms were used in the three models for 1998-2017 data and 2010-2017 data.

1)Logistic Regression

2)Random Forest

1998-2017 data:

Model 1: Basic match features like Day Number, Score difference, Seed Difference

Model 2: Past performance of the teams in different year ranges ***Model 3:***

Combination of 1 and 2 with Cross validator **2010-2017 data:**

Model 1: Basic match features like Day Number, Seed Difference, past performance for different year ranges (1998-2005,2006-2010,2011-2015,2016-2017)

Model 2: Basic features with various ratios of match characteristics such as goals, three pointers, fouls, blocks, steals, Rebounds

Model 3: Combination of 1 and 2 with Cross Validator

3.3 Cross validator parameters:

For Logistic Regression- Elastic Net Regularization:0,0.5,1 number of folds:5; Regular params:0.01, 0.5, 2; Max Iterations:5

For Random Forest- Max depth: (2,4,6); Max Bins: (20,60); Number of Tress: (5,20)

The data sets for 1997-2017,2010-2017 were divided into training, validation and testing sub parts (6:3:1 ratio).

Models were fit on training dataset and transformed on validation dataset. The best model was selected based on the highest AUC value for the validation data.

The best model was used to predict the winners of each match in 2018 data set.

3.4 Feature Engineering:

The feature engineering techniques used in the project are:

- **Standard scalar** -To normalize the values such as number of goals

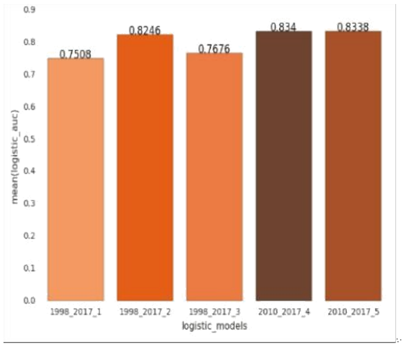
- **Quantile discretizer** -To split past performance values into multiple buckets having same number of data points
- **Bucketizer** – To split score difference values into multiple buckets based on provided splits.

4 RESULTS

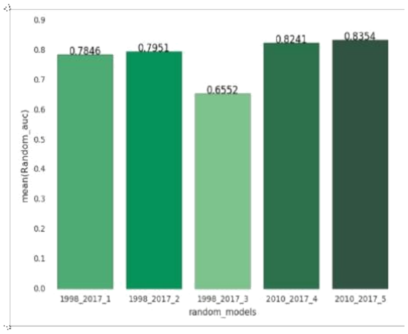
The following models were run using logistic regression and random forest algorithms:

Model Name	Logistic Regression Validation AUC	Random Forest Validation AUC	Features
1998-2017 Model 1	0.7508	0.7846	Day Number, Winning team ID, Score Difference, Location, Seed Difference
1998-2017 Model 2	0.8246	0.7951	Winning team ID, 1998-2005win percentage, 2006-2010winpercentage,2011-2015win percentage, 2016 2017win percentage
1998-2017 Model 3	0.7676	0.6552	Model 1 + Model 2 features, Winning team win percentage, losing team win percentage
2010-2017 Model 1	0.8282	0.8287	Day Number, Winning team ID, Score Difference, Location, Seed Difference, 2010-2015win percentage, 2016-2017win percentage
2010-2017 Model 2	0.834	0.8241	Model 1 features + Winning team win percentage, losing team win percentage, Win team goals ratio, Win team 3pointers ratio, Win team free-throws ratio, Win team accomplish points, Losing team goals ratio, Losing team 3pointers ratio, Losing team free-throws ratio, Losing team accomplish points
2010-2017 Model 3 (with cross validator)	0.8338	0.8354	Model 2 features with cross validator

Logistic Regression model comparison



Random Forest model comparison



The best model was found to be 2010-2017 model 3 using random forest algorithm with the highest validation AUC. Using the features from this model, we tested this model on the 2018 dataset and obtained an accuracy of 0.8016

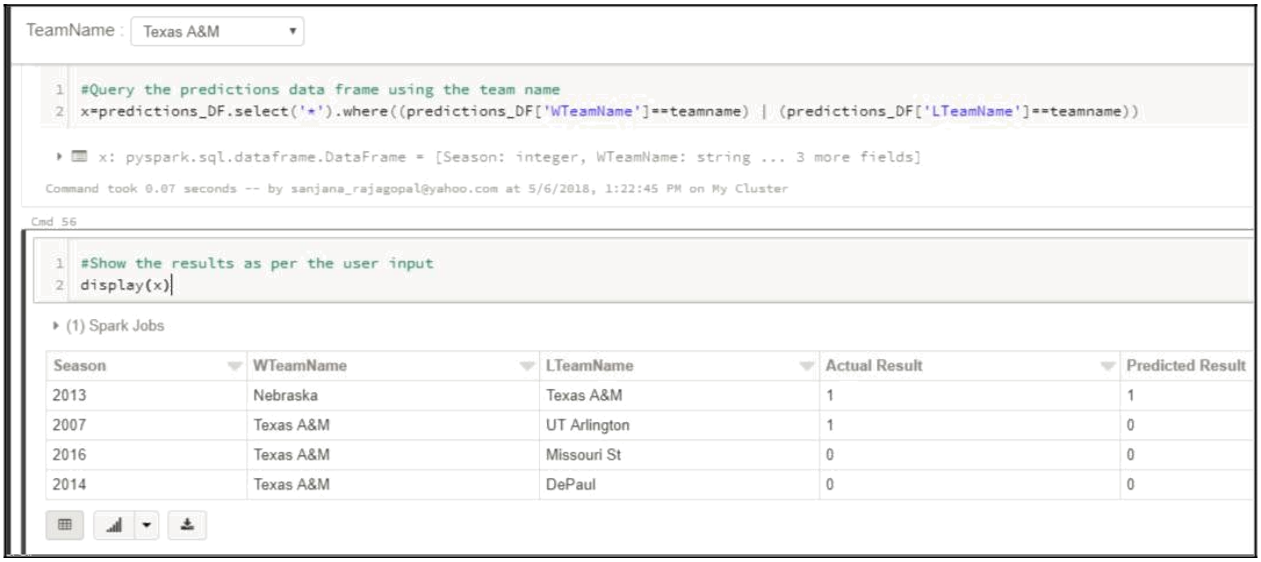
The following represent the feature importances

Features	Weights
2010 to 2015 -win percentage	0.568
Losing Team’s Goals made versus attempted ratio	0.352
2016 to 2017-win percentage	0.131
Number of Over Times	0.0903
Seed Difference	0.0101
Losing team’s combined defensive statistics	0.002
Score Difference	-0.0014
Winning team’s 3-pointers made versus attempted ratio	-0.0014
Location	-0.002
Winning team’s combined defensive statistics	-0.0038
Winning Team ID	-0.0141
Losing team’s 3-pointers made versus attempted ratio	-0.0438
Losing team’s Free throws made versus attempted ratio	-0.0549
Winning team’s overall win percentage	-0.0845
Winning team’s Goals made versus attempted ratio	-0.099
Winning team’s Freethrows made versus attempted ratio	-0.1767
Losing team’s overall win percentage	-0.3485

From the weights shown above, the most significant feature is observed to be the overall past performance of the team between the years 2010 to 2015.

4.1 User Interaction

We went beyond the project’s goals and implemented a widget to showcase a demo of the predictive model. The dropdown widget contains the list of team names. The user can choose a team’s name and the winning probability of the team against all the other possible teams in the season 2018 will be displayed.



TeamName : Texas A&M

```
#Query the predictions data frame using the team name
x=predictions_DF.select('*').where((predictions_DF['WTeamName']==teamname) | (predictions_DF['LTeamName']==teamname))
```

```
x: pyspark.sql.dataframe.DataFrame = [Season: integer, WTeamName: string ... 3 more fields]
Command took 0.07 seconds -- by sanjana_rajagopal@yahoo.com at 5/6/2018, 1:22:45 PM on My Cluster
```

Cmd 56

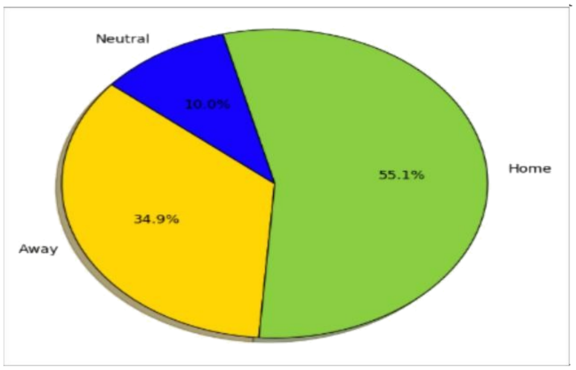
```
#Show the results as per the user input
display(x)
```

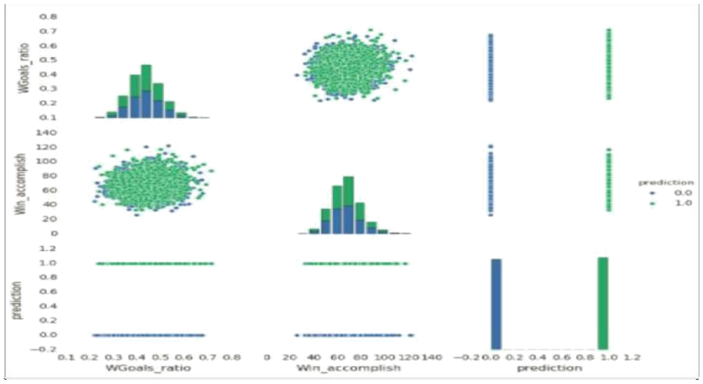
(1) Spark Jobs

Season	WTeamName	LTeamName	Actual Result	Predicted Result
2013	Nebraska	Texas A&M	1	1
2007	Texas A&M	UT Arlington	1	0
2016	Texas A&M	Missouri St	0	0
2014	Texas A&M	DePaul	0	0

4.2 Inference

From the win predictions obtained, it was observed that 55% of the teams won the games on home ground, 34.9 % won outside home and 10% won on a neutral ground. From the pair-plot, it is evident that goals ratio directly affects the winning probability of a team.





5 DISCUSSION AND CONCLUSION

While the foundation step in this project was application of statistical and machine learning concepts for development of a predictive model, the interesting step was exploring beyond and connecting our findings. These are the non-obvious, compelling observations –

- The past performance of the individual teams spread across different year ranges- 1998-2005, 2006-2010, 2011-2015, 2016-2017 has a significant impact on the prediction; the performance in 2011-2015 was more influential when compared to that of the previous year 2016-17.
- The difference in the match characteristics specific to the goals, fouls and 3-pointers of the two teams was found to positively affect the winning probability of the teams in a match game.
- Considering the higher percentage of the teams that won on home ground than those played outside, it can be concluded that the chance of win varies directly with the venue of the match.

As part of the future enhancements, we intend to experiment with other classification algorithms and obtain improvement in the model performance. Additionally, the capability of the model can be extended to predict the overall winning team of the tournament. This serves as a motivation for taking this project forward and discovering new insights from the data.

6 REFERENCES

Link to data source: https://www.kaggle.com/c/womens-machine-learning-competition-2018/data

https://en.wikipedia.org/wiki/NCAA_Division_I_Women%27s_Basketball_Tournament

Apache spark-MLlib documentation: https://spark.apache.org/mllib/

https://www.geekwire.com/2016/secrets-sports-analytics-top-athletes-coaches-execs-explain-importance-analyzing-data/
https://www.sciencedirect.com/science/article/pii/S2210832717301485

https://www.datanami.com/2017/05/26/deep-learning-revolutionize-sports-analytics-heres/

https://fivethirtyeight.com/features/how-fivethirtyeight-is-forecasting-the-2017-ncaa-tournament/

https://weiminwang.blog/2016/06/09/pyspark-tutorial-building-a-random-forest-binary-classifier-on-unbalanced-dataset/

LINK TO CODE REPOSITORY

https://github.com/Sanjana-Rajagopala/NCAA_Predictions_2018