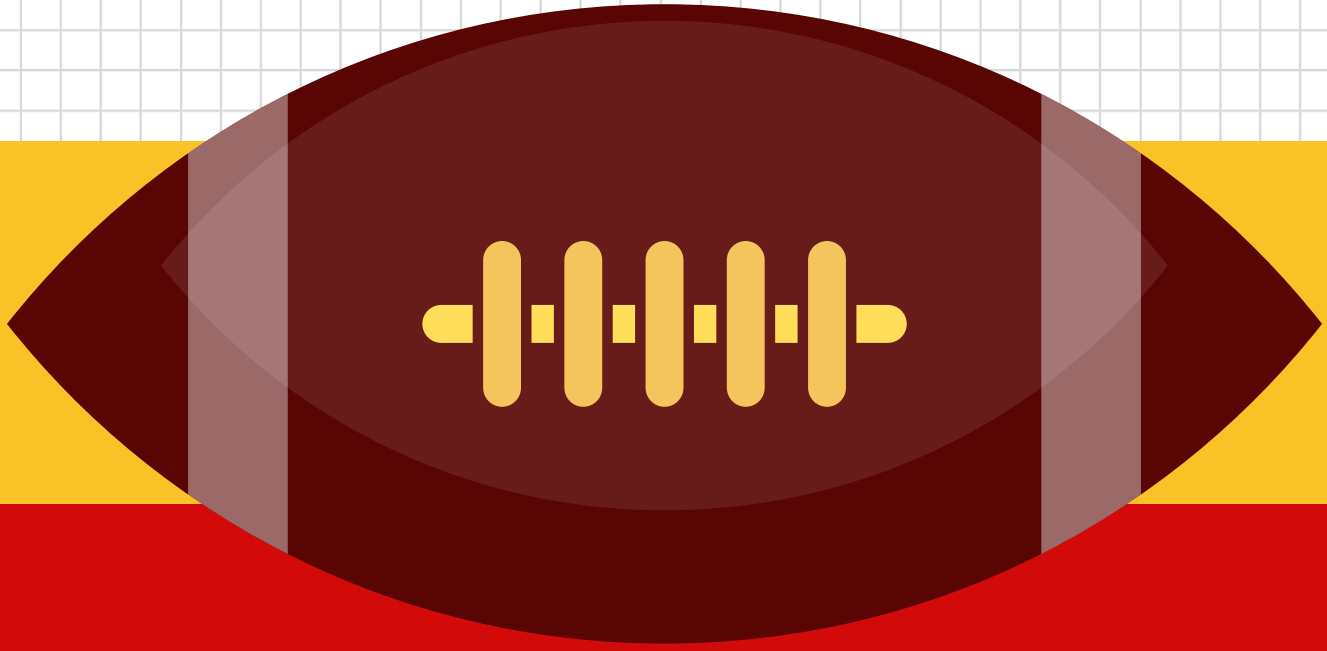
BANA 290: FORECASTING

SUPERBOWL ASSIGNMENT



ANKIT JAIN

ID: 57806263

OBJECTIVE

The Superbowl is the annual championship game of the National Football League (NFL), and it has been played every first Sunday of February every year since 2004. It is essentially the final match for the football season which begins during the late summer of the previous year. This year, the 55th Superbowl match is scheduled to take place on Feb 7th, 2021 at the Raymond James Stadium in Tampa, Florida. The match is going to be between the Tampa Bay Buccaneers and the Kansas City Chiefs.

VS

The objective of the assignment is to create a predictive model to predict the scores and winner of the Superbowl Match.

DATA SOURCES

Dataset: 2020-21 Season Matches Statistics

This dataset encompassed statistics such as team scores, number of passing and rushing yards for both teams, number of turnovers per team for every single match that has taken place during the course of this NFL season.
Source: https://www.pro-football-reference.com/

Dataset: Touchdown Log Data

This data included the number of touchdowns made by either team during each match of the season.
Source: https://www.pro-football-reference.com/

Dataset: ELO Data about each team

This dataset included the adjusted ELO score for each team based on their previous season performances and adjusted for quarterback experience as well. It also included the probability of the team winning based on their ELO scores.
Source: https://fivethirtyeight.com/

DATA PRE-PROCESSING

For this project, I used Python on Jupyter Notebooks to create my model. These were the following steps I took to process my data:

1. I cleaned the dataset such that it has only one instance of each match.
2. I added a binary for the location (1 = Home, 0 = Away) and also a column to state who the winner and non-winner of the match was.
3. I also created another dataset that encompassed the statistics for all 32 teams, before the Superbowl, such as average score, average passing and rushing yards, the average number of turnovers, number of wins and losses, number of games played at home and away, current ELO score etc.

```
teams_df.head()
```

	Team	Number of Matches	Home Games	Away Games	Away Wins	Home Wins	Away Losses	Home Losses	Total Wins	Total Losses	...	Max Rushing	Max Turnovers	Min Score	Min Passing	Min Rushing	Min Turnovers	Ac
0	Cleveland Browns	18	8	10	6	6	4	2	12	6	...	307.0	3	6.0	122.0	45.0	0	
1	Baltimore Ravens	18	8	10	7	5	3	3	12	6	...	404.0	4	3.0	70.0	110.0	0	
2	Buffalo Bills	19	10	9	6	9	3	1	15	4	...	190.0	3	16.0	122.0	32.0	0	
3	Philadelphia Eagles	16	8	8	1	3	7	4	4	11	...	246.0	4	14.0	98.0	57.0	0	
4	Denver Broncos	16	8	8	3	2	5	6	5	...	...	189.0	5	...	2.0	...	0	

MODEL SELECTION

Based on research and previous understanding of predictive models, before making a decision on which model to pursue, I wanted to try out different modelling techniques.

- Technique 1 - Clustering: I created clusters based on the overall team statistics from the dataset metrics I created, and I tried from 2 to 10 clusters numbers. Based on Euclidean Distance as a similarity measure, the two teams playing in the Superbowl this year are similar so they ended up in the same cluster always, so this method wasn't ideal to predict the score.
- Technique 2 - Regression: I attempted to predict the score linearly using rolling statistics based on when in time the match took place. Linear regression gave results but some of them were negative and sometimes the intercept didn't make sense.

I also attempted ARIMA and ETS models on Alteryx and Excel, but the regression model seemed to have a more realistic score predictions, so I decided to pursue that more but also looked into changing the degree of variables, and go for a polynomial regression approach.

MODEL DESCRIPTION

I created a function on Python which created a model for each individual team based on their respective statistics, but using the same variables, so the base model is the same. I do an 80-20 percentage split in terms of the matches played to train and test the model. Based on which score predicted for either team is higher, I decide the winner based on that.

First, I checked which degree of different variables seems to have the most linear relationship with the score of the team. The statistics of the teams are based on what the team had achieved till that particular match. For example, if we are predicting the score of the Carolina Panthers in their match against the Arizona Cardinals, which was their 4th match respectively in the season, we would consider the statistics gained after the first 3 matches both the teams have played.

After seeing which degree might be best for different variables, I tried out different combinations of both exponent degrees and variables and came up with the case which seemed most realistic with the best accuracy.

MODEL

Consider there are two teams playing in a match: Team 1 and Team 2, we are trying to predict the score of Team 1 in the match to be played.

Variables Considered

- Score: Average Score of Team 1 based on previous matches played by them in the season
- Passing: Average Number of Passing Yards of Team 1 based on previous matches played by them in the season
- Rushing: Average Number of Rushing Yards of Team 1 based on previous matches played by them in the season
- Turnover: Average Number of Turnovers of Team 1 based on previous matches played by them in the season. The degree of this variable is -1.
- Location Advantage: Probability of Team 1 winning a match given that it is being played at a particular location based on their previous matches in the season.
- Turnover Range: Difference between the maximum and minimum number of Turnovers of Team 1 based on matches already played so far. The degree of this variable is -1.
- Average Touchdowns: Average number of touchdowns made by Team 1 in the season so far. The degree of this variable is -25.
- Score Difference: Difference between the ELO Scores of Team 1 and Team 2 just before the match.
- Opponent Score: Average Score of Team 2 based on previous matches played by them in the season
- Opponent Passing: Average Number of Passing Yards of Team 2 based on previous matches played by them in the season
- Opponent Rushing: Average Number of Rushing Yards of Team 2 based on previous matches played by them in the season
- Opponent Turnover: Average Number of Turnovers of Team 2 based on previous matches played by them in the season. The degree of this variable is -1.

MODEL ACCURACY METRICS

Kansas City Chiefs

- R2=0.999
- MAPE = 2.23%

Tampa Bay Buccaneers

- R2=0.7507
- MAPE = 10.80%

MODEL PREDICTION

24	Tampa Bay Buccaneers
32	Kansas City Chiefs

Prediction: Kansas City Chiefs will win Superbowl 55 against Tampa Bay Buccaneers with a score of 32 to 24.

MODEL LIMITATIONS AND FUTURE STEPS

- The current model is based on the assumption that teams have at least played one match, and it will not work to predict the score of the team if they haven't played at least one match before, so I would work on making the model more applicable to matches with no previous statistics.
- The model seems to be more working better in predicting the score for the Chiefs as opposed to the Buccaneers from the accuracy metrics, I believe that I need to work on making the model creation more accurate for either team.
- There are a lot of variables that can be considered such as player statistics, coach statistics, weather at the time of the match, number of MVP titleholders playing, so for the future, I can work on adding those to the dataset.
- There is a possibility of overfitting of the model based on the 80%-20% split, so for future steps, I would work on making the testing set bigger as well to make a more informed and accurate model.

ENSEMBLE MODELLING

- Team: Ankit Jain, Marianna Carini, Karl Hickel, Allen Lee, Mira Daya
- My team and I each made one or more models to predict the score for the Superbowl and then took a weighted average based on the value of 1 - MAPE to get the score for both teams.
- We used Excel to create the weights and obtain the average.

Model	Source	KC	Tampa	MAPE	Weight KC	Weight TB
Lazy Man's (Avg.)	Marianna	32	26	0.17	0.12	0.13
Lazy Man's (MAPE Avg.)	Marianna	31	26	0.17	0.12	0.13
Success Rate Only	Marianna	29		0.18	0.12	0.00
Success Rate Only	Marianna		34	0.29		0.11
Success Rate + In-game Stats (Avg.)	Marianna	22		0.14	0.12	0.00
Success Rate + In-game Stats (Avg.)	Marianna		30	0.24		0.12
Poly. Reg	Ankit	32		0.02	0.14	
Poly. Reg	Ankit		24	0.11		0.14
Lin. Reg	Mira	32		0.06	0.13	
Lin. Reg	Mira		26	0.23		0.12
Lin. Reg	Karl	31		0.15	0.12	
Lin. Reg	Karl		29	0.25		0.12
Lin. Reg	Allen	30		0.07	0.13	
Lin. Reg	Allen		28	0.20		0.13

TEAM ENSEMBLE MODEL PREDICTION

28	Tampa Bay Buccaneers
30	Kansas City Chiefs

Prediction: Kansas City Chiefs will win Superbowl 55 against Tampa Bay Buccaneers with a score of 30 to 28.