Análisis de Desempeño de Equipos de la NFL y predicciones de la temporada regular 2019

NFL Team Performance Analysis and Predictions for the NFL 2019-2020 season

Dieter Esteban de Wit y Antonio Reyes

28 de mayo de 2020

Resumen

El presente informe es un resumen del trabajo realizado como proyecto final del curso de Minería de Datos para la carrera de Ciencias de la Computacion y Tecnologias de la Información de la Universidad del Valle de Guatemala por los alumnos Antonio Reyes y Dieter de Wit. En el presente trabajo intentamos realizar la predicción de la cantidad de partidos que ganaría cada equipo en la temporada regular de la NFL (National Football League) basado en los datos de la ofensiva de los 32 equipos en las temporadas de 2014 al 2018. Se utilizaron los algoritmos K-Means Clustering para la división de los datos numéricos en data categórica, Redes Neuronales basados en una arquitectura de 3 capas, 32 Neuronas por capa, con modelo Secuencial basado en Stochastic Gradient Descent Optimizer para poder predecir todos los datos de ofensiva, y por último, se utilizó Regresión Lineal para poder predecir la cantidad de partidos ganados por cada uno de los equipos. Se pudo concluir que el proyecto tuvo éxito en su área, no un éxito de precisión pero sí un éxito comparativo realizado con una predicción llamada “For the Win” donde nuestro modelo logró datos más acertados.

Abstract

This report is a summary of the work carried out as a final project of the Data Mining course for the Computer Science and Information Technology career at "Universidad del Valle de Guatemala" by the students Antonio Reyes and Dieter de Wit. In the present work we try to make a prediction for the number of games won by each team in the 100th season of the NFL (National Football League) based on the offensive data of the 32 teams in the seasons from 2014 to 2018. The K-Means Clustering Algorithm was used to divide numerical data into categorical data, a Neural Network based on a 3-layer architecture, 32 Neurons per layer, with a Sequential model based on Stochastic Gradient Descent Optimizer was used to predict all the offensive data, and lastly, Linear Regression was used to predict the number of games won by each of the teams. It was concluded that the project was successful in its area, not a precision success but a comparative success made with a prediction called “For the Win” where our model achieved more accurate data.

Introducción

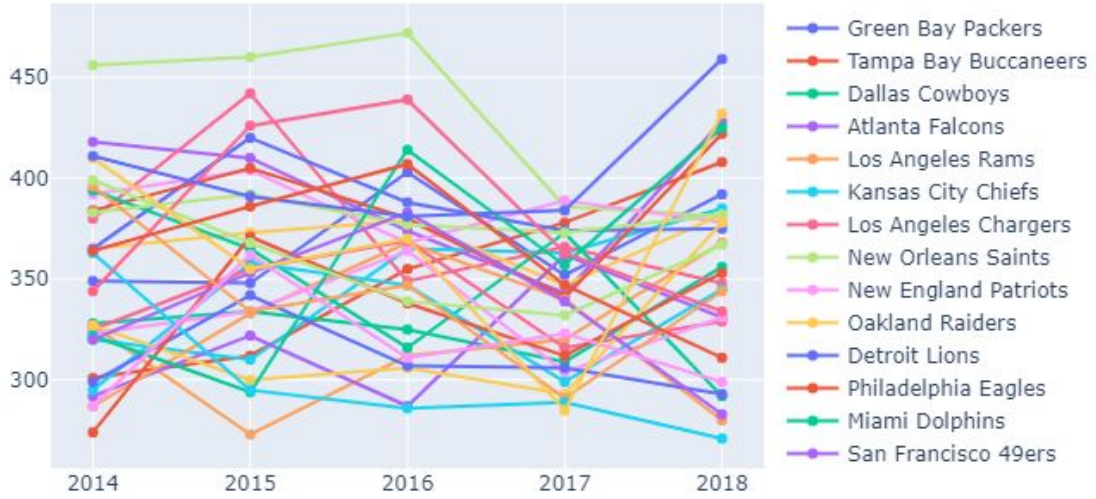
"La NFL (National Football League) generó un aproximado de $15 billones la temporada 2018, lo cual los corona como los reyes de los deportes en los Estados Unidos." - *Chicago Tribune*

El football americano genera miles de datos sobre cada uno de los jugadores y sobre el equipo de general por cada partido, día de entrenamiento, lugar donde se realizará el partido, etc. Este deporte y el Baseball podrían considerarse los deportes que recurren más al análisis de datos para elegir a sus jugadores y para evaluar el rendimiento del equipo a lo largo de la temporada.

El desempeño de un equipo de fútbol americano puede puede cambiar drásticamente con el más mínimo cambio (o en alguno casos por no querer realizar cambios), esta irregularidad en el desempeño de equipos se representa en **Gráfica 1** que muestra los pases totales realizados por cada equipo de la NFL entre las temporadas 2014-2018. Numerosas organizaciones tratan cada año de predecir recaudar la mayor cantidad de información posible para tratar de predecir los resultados de la siguiente temporada, ya sea por uso de algoritmos y análisis de datos o por opiniones de expertos del deporte o ex jugadores.

En base a la información obtenida de previas temporadas, se evaluará el desempeño de la ofensiva de cada uno de los equipos para categorización de equipos Sobresalientes, Aceptable y Abajo del Promedio para determinar y por medio de algoritmos de aprendizaje, se determinará si los cambios realizados por dichos equipos se reflejan en la última temporada (2019), También se utilizarán dichos algoritmos de aprendizaje para proveer una predicción de la cantidad de victorias de los equipos en la temporada 2019 en base a su desempeño de temporadas previas. Esta predicción se comparará con la predicción realizada por la página de deportes "For the Win".

Gráfica 1: Pases totales realizado por cada equipo entras las temporadas 2014-2018.


450
400
350
300
2014
2015
2016
2017
2018
Green Bay Packers
Tampa Bay Buccaneers
Dallas Cowboys
Atlanta Falcons
Los Angeles Rams
Kansas City Chiefs
Los Angeles Chargers
New Orleans Saints
New England Patriots
Oakland Raiders
Detroit Lions
Philadelphia Eagles
Miami Dolphins
San Francisco 49ers

Materiales y métodos

La información de la ofensiva de los equipos y el número de victorias en las temporadas fueron obtenidas de la página "Football Database".

K-Means Clustering

Se utilizó la información de anotaciones por aire y por tierra realizadas por cada equipo en las temporadas de 2014 al 2018 para clasificar a los equipos en base a 3 grupos de datos que tomarían en cuenta los el total de anotaciones de los equipos por medio de pase y acarreos, se consideran los grupos como: Outstanding, Above Average, Below Average. Se conoce como dato de la NFL que la cantidad de anotaciones y yardas obtenidas por tierra o por aire genera un balance a la ofensiva de cada equipo, ya sea sesgado a alguna de las dos, sobresaliente o con carencias en ambas.

Redes neuronales

Esta agrupación también realizó con la información de temporadas previas, esta data sirvió como data de entrenamiento. Por medio de redes neuronales, se realizó una predicción de todos los valores de ofensiva para los 32 equipos en la temporada regular del 2019. Los datos ingresados a la Red Neuronal que fueron relevantes son: Anotaciones por tierra, Anotaciones por acarreo, Retornos de patada, Retornos de despeje, Intercepciones y Goles de Campo. Se obtuvo una precisión de más del 80% en esta Red Neural para casar con los datos de la temporada regular del 2019.

Regresión lineal

Utilizando los resultados de las 4 temporadas previas, se utilizó Regresión lineal para predecir el posible resultado que represente la cantidad de victorias de la temporada regular 2019 para cada equipo, este resultado fue comparado con el valor real y se clasificaron los resultados en "Diferencia 0" (cuando el resultado predicho era igual al resultado real), "Diferencia 1-3" y "Diferencia mayor a 3". Se realizó lo mismo con la predicción de la página "For the Win" (que se muestra en **Imagen 1**) y se compararon los resultados.

Imagen 1: Predicción realizada por "For the Win".

PROJECTED 2019 NFL STANDINGS

AFC

EAST	NORTH	SOUTH	WEST
TEAM	**TEAM**	**TEAM**	**TEAM**
1. Patriots (13-3)	1. Browns (10-6)	1. Colts (12-4)	1. Chargers (13-3)
2. Bills (6-10)	2. Ravens (10-6)	2. Texans (6-10)	2. Chiefs (11-5)
3. Jets (4-12)	3. Steelers (9-7)	3. Titans (5-11)	3. Broncos (8-8)
4. Dolphins (4-12)	4. Bengals (4-12)	4. Jaguars (4-12)	4. Raiders (6-10)

NFC

EAST	NORTH	SOUTH	WEST
TEAM	**TEAM**	**TEAM**	**TEAM**
1. Eagles (11-5)	1. Vikings (11-5)	1. Saints (13-3)	1. Rams (13-3)
2. Cowboys (8-8)	2. Packers (10-6)	2. Falcons (11-5)	2. Seahawks (9-7)
3. Redskins (7-9)	3. Bears (8-8)	3. Buccaneers (8-8)	3. 49ers (4-12)
4. Giants (4-12)	4. Lions (5-11)	4. Panthers (6-10)	4. Cardinals (4-12)

Resultados y discusión

K-Means Clustering

La **Gráfica 2** muestra la agrupación de los equipos de la temporada 2019 y la **Imagen 2** muestra los equipos que pertenecen a cada categoría. En base a los resultado de la gráfica, puede verse claramente que lo que define la clasificación al grupo "Below Average" es a los equipos cuya cantidad de anotaciones por acarreos es casi inexistente y su cantidad de anotaciones por pase largo es menor al promedio. Los equipos que pertenecen al grupo "Above Average" poseen valores más balanceados entre la cantidad de anotaciones por pase y por acarreo. Finalmente, los equipos pertenecientes al grupo "Outstanding" poseen un balance en la cantidad de anotaciones por pase largo y por acarreo y su total excede por mucho a lo equipos de las otras categorías. Este método nos proporcionó la facilidad de poder traducir variables numéricas cuantitativas a variables cualitativas para poder realizar análisis posteriores.

Gráfica 2: Agrupación de equipos de la temporada regular 2019

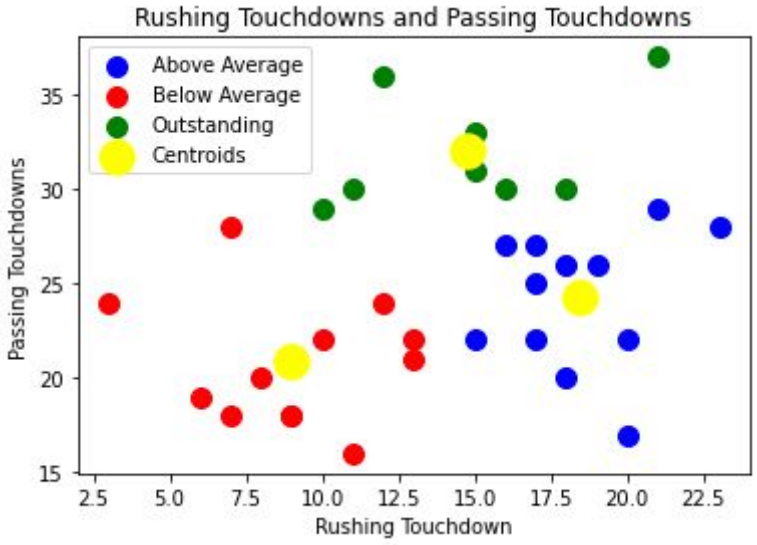


Imagen 2: Equipos que pertenecen a cada categoría en los Clusters.

BELOW AVERAGE

- Jacksonville Jaguars
- Detroit Lions
- New York Jets
- Pittsburgh Steelers
- Chicago Bears
- Washington Redskins
- Denver Broncos
- Miami Dolphins
- Los Angeles Chargers
- Oakland Raiders
- Buffalo Bills
- Cincinnati Bengals

ABOVE AVERAGE

- Cleveland Browns
- Philadelphia Eagles
- Houston Texans
- Green Bay Packers
- Minnesota Vikings
- Tennessee Titans
- San Francisco 49ers
- New England Patriots
- Indianapolis Colts
- Los Angeles Rams
- Arizona Cardinals
- Carolina Panthers

OUTSTANDING

- Atlanta Falcons
- New York Giants
- Dallas Cowboys
- Kansas City Chiefs
- Seattle Seahawks
- Tampa Bay Buccaneers
- New Orleans Saints
- Baltimore Ravens

Redes Neuronales

Se realizó la agrupación de equipos con información de temporadas previas, estos resultado fueron utilizadas información de entrenamiento de los datos para poder predecir las categorías a las que pertenecen los equipos. Utilizando una red con 32 neuronas y 1000 iteraciones, se obtuvo más de 80% de certeza. Este resultado se debe a que la agrupación de los equipos se realiza en base a el desempeño de dicha temporada y no en las temporadas como un colectivo. Si hubo una temporada en la que todos los equipos tuvieron un mejor desempeño en base a su temporada anterior, se aumenta la posibilidad de error. **Imagen 3** y **Gráficas 3** presentan los resultados. También, la Red analiza los datos de ofensiva: Retornos de patada, Retornos de despeje, Intercepciones y Goles de Campo de las temporadas 2014 a 2018 y predice cuáles serán los datos finales para la temporada regular 2019.

Imagen 3: Porcentaje de certeza del modelo de Redes Neuronales.

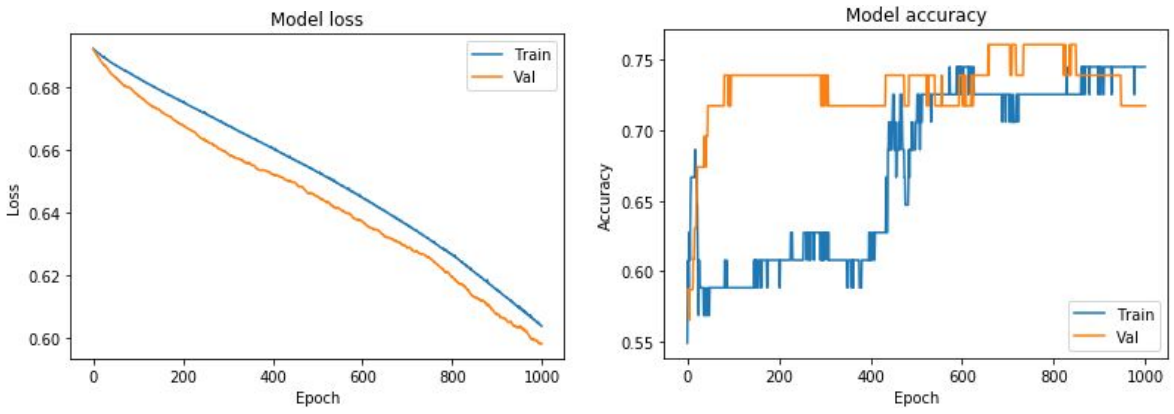
Certeza

```
model.evaluate(X_test, Y_test)[1]
```

```
31/31 [==============================] - 0s 37us/step
```

```
0.8064516186714172
```

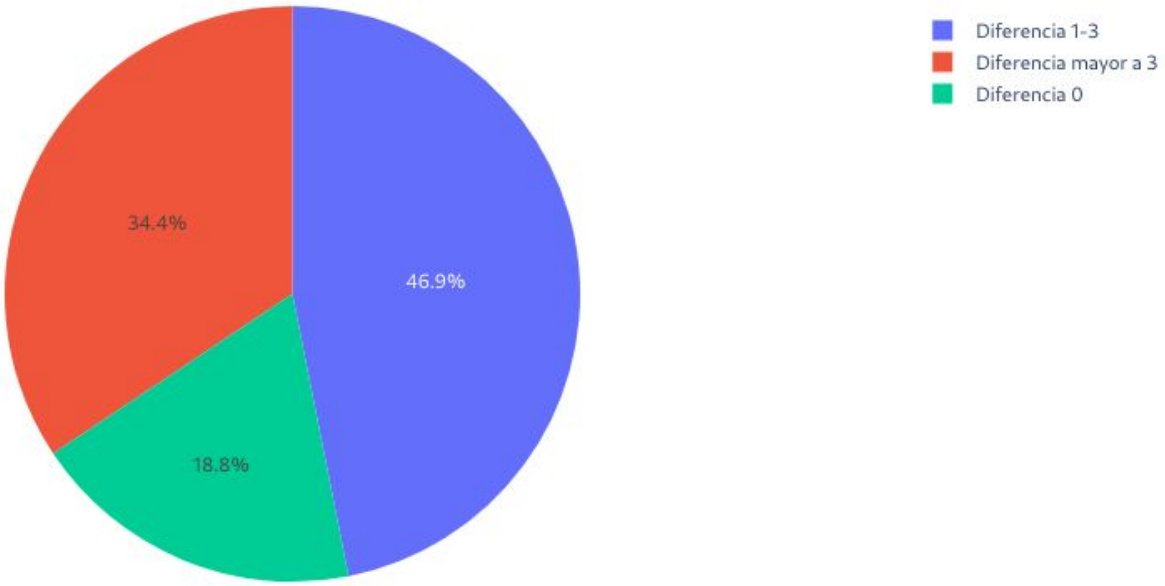
Gráfica 3:



Regresión logística

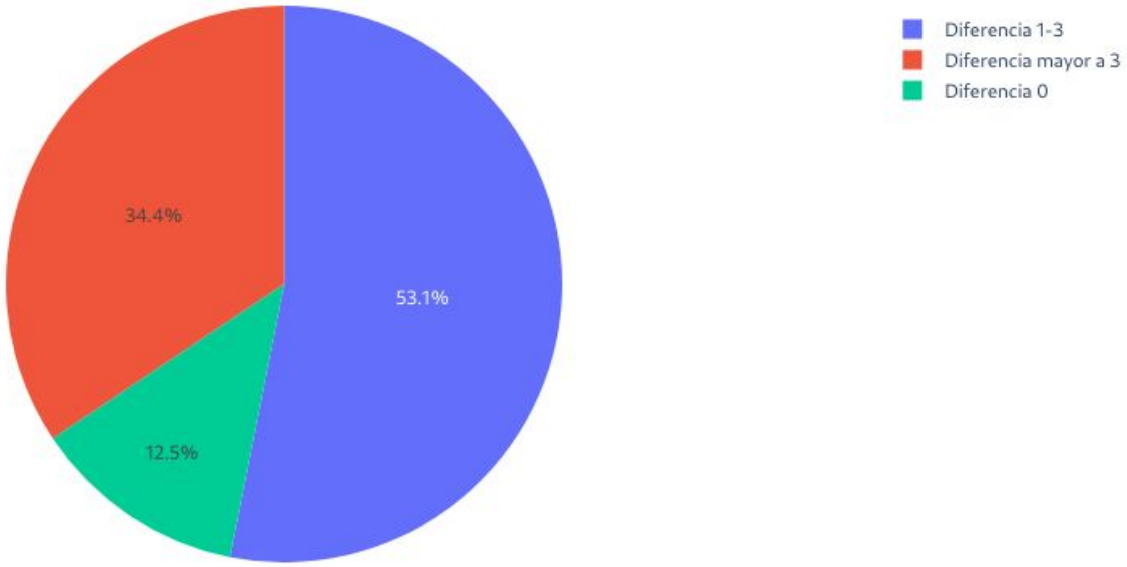
Los resultados de **Gráfica 5** muestran que se logró acertar la cantidad de victorias exactas del 18.8% de los equipos y un 46.9% con valores cercanos al real. El significante porcentaje de en los equipos cuyos puntajes predichos tienen un margen de error mayor a 3 puntos se debe a que se tenían pocos datos de cada equipo, sólo se tenían 4 valores que correspondían a las 4 temporadas pasadas, un mal año o un año excepcional (a comparación del resto) afectaron de manera significante el resultado final.

Gráfica 5: El Pie Chart presenta el porcentaje de predicciones cada categoría en la predicción utilizando Regresión lineal



A pesar que el porcentaje de predicciones erróneas tenía un tamaño significante, estos resultados se consideran más confiables que las predicciones realizadas por “For the Win” que se muestran en **Gráfica 6**.

Gráfica 6: El Pie Chart presenta el porcentaje de predicciones cada categoría en la predicción realizada por “For the Win”



Conclusiones

La clasificación de equipos en base a su desempeño de su ofensiva en la temporada probó ser efectiva ya que crea una clara separación en base al los valores de anotaciones en pase a distancia y acarreos y la predicción de dicha clasificación en la temporada 2019 dió valores y un porcentaje considerablemente válido ya que este es un deporte que cuenta con una complejidad y cantidad de datos mucho mayor a cualquier otro.

La predicción victorias a lo largo de la temporada por medio del uso del algoritmo de regresión lineal mostró ser más confiable que la predicción realizada por "For the Win" a pesar de tener un porcentaje significante de error, esta predicción puede mejorarse aumentando la cantidad de información que se provee al algoritmo obteniendo los resultados de más temporadas. También se propone la mejor forma de realizar este análisis con un porcentaje de confiabilidad aun mucho mayor sería el realizar la red neuronal sobre los datos de defensiva, preparar una comparación entre defensiva y ofensiva por equipo a enfrentarse en la temporada y luego realizar la regresión que competa (lineal o polinomial). En resumen, el proyecto, con cantidad de tiempo considerable, puede ser reproducido y mejorado para un análisis que puede ser muy significativo en el mercado, considerando el potencial que tuvo sobre el análisis realizado por USA Today.

Bibliografía

Ruiz, S. (3 de Mayo de 2019). *For the Win*. Obtenido de https://ftw.usatoday.com/2019/05/2019-nfl-season-predictions-every-game-super-bowl

The Football Database. (2020). Obtenido de https://www.footballdb.com/index.html