Informe Avance Predicción de Ubicación Final En Una Partida de PUBG

Por:

Jhon Alexander Bedoya Carvajal

Maria Camila Arcila Ramírez

Curso:

Introducción a la Inteligencia Artificial

Docente:

Raúl Ramos Pollan

Universidad de Antioquia

Facultad de Ingeniería

Medellín

2022

Contenido

1. Introducción

Uno de los pilares más importantes para la Inteligencia Artificial, es precisamente el poder analizar como nosotros los seres humanos podemos buscar distintas soluciones para problemas que se nos presentan día con día en todo lugar y aún más importante el saber cómo somos capaces de poder seleccionar una solución entre tantas posibles y poder resolver los imprevistos que se nos presentan. En este trabajo se tiene a disposición más de 65,000 juegos en datos de jugadores anónimos, divididos en conjuntos de entrenamiento y prueba, y se busca predecir la posición en el ranking final a partir de las estadísticas finales del juego y las calificaciones iniciales de los jugadores.

2. Planteamiento el problema

Los videojuegos estilo Battle Royale han conquistado el mundo. 100 jugadores caen en una isla con las manos vacías y deben explorar, buscar y eliminar a otros jugadores hasta que solo quede uno en pie, todo mientras la zona de juego continúa reduciéndose. PlayerUnknown's BattleGrounds (PUBG) ha disfrutado de una gran popularidad. Con más de 50 millones de copias vendidas, es el quinto juego más vendido de todos los tiempos y tiene millones de jugadores activos mensuales. Se busca predecir la posición en el ranking final a partir de las estadísticas finales del juego y las calificaciones iniciales de los jugadores. ¿Cuál es la mejor estrategia para ganar en PUBG? ¿Deberías sentarte en un lugar y esconderte en tu camino hacia la victoria, o necesitas ser el mejor tirador? ¡Dejemos que los datos hablen!

2.1. Dataset

El Dataset proviene de una competencia de Kaggle en la que se proporciona datos de más de 65 mil partidas de PUBG. Este Dataset cuenta con 29 columnas y 4446966 filas.

El archivo de entrenamiento tiene el nombre de “train_V2.csv”, el de prueba tiene el nombre de “test_V2.csv” y el archivo que contiene un ejemplo de envío es llamado “sample_submission_V2.csv”.

Debido a la cantidad tan extensa de datos, se tomó una muestra aleatoria de 200.000 filas para trabajar en este proyecto. Inicialmente, se buscó cumplir con los requisitos establecidos para el Dataset (al menos el 10% de las columnas han de ser categóricas y al menos ha de tener un 5% de datos faltantes en al menos 3 columnas). Las columnas 'killPlace', 'maxPlace' y 'matchType' se transformaron a categóricas. Además, como lo explica la descripción del Dataset, las columnas ‘rankPoints’, ‘killPoints’ y ‘winPoints’ cuentan con valores que deben ser tratados como variables categóricas. Por último, se eligieron 3 columnas aleatoriamente para eliminar entre el 5% y 10% de las entradas para cumplir con el requisito de los datos faltantes.

2.2. Métrica

La métrica de evaluación para esta competencia es el error absoluto medio (MAE) entre el winPlacePerc predicho y el winPlacePerc observado. El MAE es una medida de errores entre observaciones emparejadas, se calcula como:

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} = \frac{\sum_{i=1}^{n}|e_i|}{n}$$

Es un promedio aritmético de los errores absolutos $|e_i| = |y_i - x_i|$, donde y_i es la predicción y x_i el verdadero valor.
En cuanto a la métrica de negocio, nuestro modelo de predicción de la ubicación final del jugador debería tener un MAE <=10% ya que se usará el modelo para determinar cuál es la mejor estrategia para ganar el juego, lo que se busca es mediante las ubicaciones de los jugadores ver si se tiene más éxito cuando el jugador cae en un lugar desolado y se esconde, o si es mejor estrategia salir a competir con los demás jugadores, por lo que es importante tener el MAE por debajo del 10%

2.3. Variable Objetivo

La variable que se quiere predecir es la posición final a partir de las estadísticas finales del juego y las calificaciones iniciales de los jugadores. A partir de esto, determinar cuál puede ser la mejor estrategia para ganar una partida de PUBG.

3. Exploración descriptiva del Dataset

La información en el Dataset ya se encuentra organizada, por lo que, en primer lugar, se analizan variables que a simple vista se consideren de interés:

3.1. Análisis de variable objetivo.

Para el análisis de la variable objetivo se graficó la distribución de esta y frente a otras variables de interés como las muertes y las posiciones en los diferentes rankings.

3.2. Muertes.
3.3. Daño infligido.
3.4. Racha máxima.
3.5. Muertes.
3.6. Distancia caminada.

3.7. Datos faltantes

Antes de iniciar a iterar modelos, es importante revisar los datos faltantes y decidir qué hacer con estos. En primera instancia, se revisa la cantidad de datos faltantes en el Dataset con los requisitos simulados, el resultado se muestra en la Figura 1.

```
damageDealt      10600
killPoints      119681
longestKill      14599
rankPoints       76212
walkDistance     21600
winPoints       119681
dtype: int64
```

Figura 1. Datos faltantes

3.8. Correlación de variables

En un solo gráfico, se mostraron las distribuciones de las variables de interés para analizar una frente a la otra y la relación entre ellas. Se encontraron algunas relaciones que ya se esperaban y otra que no, por lo que es información que aporta al desarrollo del proyecto. Por ejemplo, la correlación entre las muertes y la variable a predecir es directamente proporcional.

4. Tratamiento de datos

Para el tratamiento de datos se hizo un análisis de qué método de reemplazo (remplazar por cero, reemplazar por la media, reemplazar por valores de una normal equivalente) es el que mejora más significativamente el modelo. Con una prueba de hipótesis e iteración de varios modelos como Arboles de decisión, de encontró que, dependiendo de la variable, un método si puede aportar mejoría al Dataset mientras que otro no. Por ejemplo, para la variable "walkDistance", los tres métodos de reemplazo son significativos.

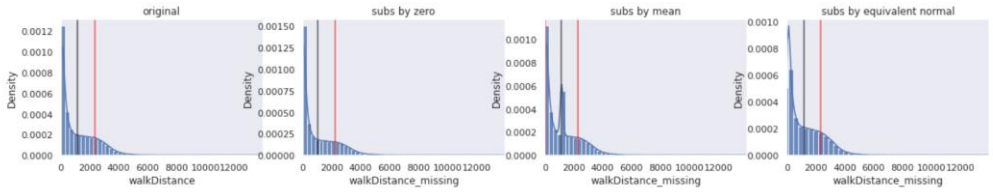

Figura 2. Métodos de reemplazo de valores faltantes en "walkDistance".

```
100% (4 of 4) |#########################| Elapsed Time: 0:04:15 Time:  0:04:15
Ttest_indResult(statistic=275.43259707307203, pvalue=3.3801071096642944e-17)
Ttest_indResult(statistic=237.3598560236848, pvalue=1.1110609257239902e-16)
Ttest_indResult(statistic=294.34443376163557, pvalue=1.987133493820836e-17)
```

Figura 3. P-values.

En resumen, hasta este punto hemos avanzado el proyecto, debido a que la extensión del Dataset elegido nos estaba dando muchos problemas para manipularlo. Este problema se solucionó tomando una muestra que permitiera ejecutar los algoritmos en memoria en Colab.