Informe Parcial #1: Análisis Exploratorio de Datos

Descripción

El conjunto de datos corresponde a una base de datos preprocesada, lista para análisis, tomada del sitio web `https://web.chessdigits.com/data`. Contiene información sobre 200k partidas de ajedrez en diferentes modalidades jugadas en `lichess.org` durante el mes de mayo de 2019. Específicamente incluye 23 variables las cuales brindan la información particular de cada una de las 200k partidas. Entre las más destacadas se encuentran el elo de cada jugador, el ECO, el resultado de la partida y los *RatingDiff* para cada jugador. Además, la base de datos contiene:

- Los primeros 100 movimientos individuales (columnas impares = blancas, columnas pares = negras)
- Tiempo restante después de cada movimiento (primeros 100 movimientos)(columnas impares = blancas, columnas pares = negras)
- Evaluación por computadora después de cada movimiento (primeros 100 movimientos)(columnas impares = blancas, columnas pares = negras)

Las variables usadas en este análisis exploratorio de datos fueron: elo, ECO, WhiteRatingDiff, número total de movimientos, resultado y categoría.

Descripción de las variables

Elo

Es la métrica canónica del ajedrez para medir el nivel de un jugador. La base de datos proporciona el elo del jugador de piezas blancas y el jugador de piezas negras.

ECO

La Enciclopedia de Aperturas de Ajedrez, por sus siglas en inglés (ECO), es un libro de referencia que describe el estado de la teoría de aperturas en el ajedrez. Se divide en 5 categorías las cuales se distinguen por las primeras 5 letras del abecedario y un número del 0 al 99.

WhiteRatingDiff

Un rating es una herramienta que facilita medir el desempeño probable de un jugador contra otro oponente. El RatingDiff es la diferencia entre el rating de las fichas blancas y negras. El valor es interpretado como la probabilidad de ganar la partida. Por lo tanto, si se asume la posición de las fichas blancas, un RatingDiff > 0 representa una mayor probabilidad de ganar la partida y un RatingDiff < 0 una probabilidad muy baja de ganar la partida.

Número total de movimientos

La base de datos contiene 200 columnas nombradas como 'Move_ply_i', donde i toma valores entre 1 y 200. Los valores impares para i representan las primeras 100 jugadas de las fichas blancas y los números pares para i dan las primeras 100 jugadas de las fichas negras.

Resultado

Contiene 4 categorías que son: 1-0, 0-1, 1/2-1/2, *. la primera categoría indica victoria para las fichas blancas, la segunda está asociada a la derrota para las fichas blancas, la tercera a tablas y la última a abandono.

Categoría

La variable categoría almacena los 4 modos de juego tradicionales del ajedrez: Bullet, Blitz, Rapid y Classic.

Plan inicial para la exploración datos

Con la base de datos seleccionada, se realizará una primera visualización de las diferentes columnas para descartar posibles variables con contenidos sin información significativa para un análisis exploratorio de datos, como la columna *Site* que incluye el URL del sitio donde se jugó la partida.

La cantidad de categorías por cada grupo y sus distribuciones serán visualizadas de forma rápida por medio de *pandas* y *seaborn* para extraer posibles relaciones entre distintas variables.

Con la primera revisión finalizada, se propondrán las primeras hipótesis acerca de la relación entre distintas variables, para posteriormente, reorganizar los datos y crear nuevas columnas con variables que faciliten las pruebas de hipótesis.

Posiblemente muchos de los datos no sean fáciles de manipular (como la columna *Opening*) o contengan valores anómalos (como las columnas *Eval_pl_i*, que son las evaluaciones de stockfish), por lo que algunas de las primeras hipótesis serán inviables y se hará una segunda revisión de la base de datos y se propondrán nuevas hipótesis teniendo en cuenta las variables que pueden ser empleadas de forma práctica.

Con las nuevas hipótesis planteadas, se modificará la base de datos para trabajar con las variables de interés, se eliminarán posibles valores que interfieran en el tratamiento de los datos y si es necesario, se crearan nuevas columnas con nuevas variables para el futuro desarrollo de las pruebas de hipótesis.

Limpieza e ingeniería de datos

En primera instancia se eliminaron las columnas con información vacía como *Round* y las columnas con contenido sin significancia como *Site*, *WhiteID* y *BlackID*. Después, una vez planteadas las hipótesis, se eliminaron las columnas que no iban a ser usadas. Estas columnas fueron principalmente datos de fechas, la hora de juego y día de la semana en la que se jugó la partida.

En la primera revisión de los datos se propuso una hipótesis que incluía el valor promedio de las 100 primeras evaluaciones hechas por stockfish para las primeras 100 jugadas hechas por el jugador blanco. No obstante, esta fue descartada debido a que las 200 columnas *Eval_pl_i* contenían una mezcla de datos de enteros y cadenas, lo que no permitió ejecutar dicho promedio. Por lo tanto, todas las columnas *Eval_pl_i* fueron descartadas. Por otro lado, las 200 columnas *Clock_ply_i* fueron eliminadas ya que no se planteó ninguna hipótesis con dicha variable.

Se creo una nueva columna llamada *EloTier* con el promedio del elo blanco y elo negro por cada partida y se dividió en cuatro categorías: Si EloTier ≥ 2000 = Very high elo, si EloTier ≥ 1800 = High elo, si EloTier ≥ 1400 = Mid elo y si no cumple ninguna de las anteriores, entonces se caracteriza por ser Low elo.

En la revisión de la base de datos se observó que la cantidad de movimientos para cada partida era diferente, es decir, muchas de las partidas terminaban antes de la jugada o movimiento número 100 o no terminaban. Por lo tanto, para las 200k partidas se decidió completar los espacios vacíos con NAN y contar el número de jugadas por partida. Con esta información se creo una nueva variable llamada *Movements*.

Por último, dado que la columna ECO contiene una gran cantidad de valores diferentes, se decidió simplificar la información categorizando cada apertura solo por su letra. De este manera se pudo hacer un conteo mucho más sencillo de las aperturas. Además, en la columna *Result* se cambiaron los nombre de las 4 categorías para facilitar la interpretación: (1-0) = Win, (0-1) = Lose, (1/2-1/2) = Draw y (*) = Abandoned.

Hipótesis

Goodness of fit para la distribución de Elo

Se usó la prueba para GOF de Kolmogorov - Smirnov (KS), sobre una muestra aleatoria de 7500 partidas con reemplazo. Este tipo de pruebas son independientes de la distribución de datos, es decir, no están ligadas a una estadística de prueba que sigue una distribución predeterminada. En esencia, es un tipo de prueba no paramétrica cuya estadística de prueba es la distancia más grande entre las gráficas de dos funciones de densidad probabilidad acumulada: una de la distribución ajustada y otra de los datos del muestreo.

- H_0: Los datos promedio de Elo tienen una distribución beta.
- H_1: Los datos promedio de Elo se desvían de una distribución beta.

En este caso se ajustaron los datos a una distribución beta con parámetros: a, entre 4 y 5; b, entre 7 y 9; loc, entre 600 y 700; y scale, entre 2400 y 2500. Claramente se trata de una distribución sesgada hacia la derecha, como el conjunto de datos; esto tiene sentido al pensar en cómo se distribuyen los jugadores: en este caso por debajo de los 2000 puntos de elo, se encuentra más del 90% de los jugadores, de forma que el ajuste debe corresponder a una distribución abultada hacia la izquierda. Este mismo tipo de distribución sesgada se puede observar en otros juegos con sistemas de clasificación similares al elo.

Ahora, teniendo en cuenta los momentos estadísticos calculados, se observa que la media es ~ 1513, con un ancho de línea de 315. Nuevamente, esto indica que el grueso de jugadores se encuentra en el intervalo $[1198, 1828]$, en específico el 66.6%. En lo que respecta a la prueba KS, se encontró un valor $p = 0.27 > \alpha = 0.05$, por la tanto la hipótesis nula no fue rechazada: los datos se ajustan a la distribución beta. Aún con esto, se debe tener en cuenta que para muestras de un tamaño mucho mayor, por ejemplo toda la población, la prueba siempre va a resultar negativa. En esos casos simplemente se tiene un caso de que cualquier desviación penaliza fuertemente al ajuste y, en consecuencia, a la estadística de prueba.

En el gráfico q-q asociado a esta prueba (ver cuaderno de Jupyter en GitHub) se observa que, salvo por los valores superiores a los 2500 de elo, los datos se ajustan muy bien a la distribución beta. La desviación en la parte superior derecha señala que ese grupo de partidas son partidas fuera de la tendencia. Como una confirmación adicional, teniendo en cuenta una división categórica del elo, se realizó una prueba gof χ^2 de carácter categórica. El test χ^2 es una comparación de la frecuencia en que resulta una muestra en una categoría con respecto a la frecuencia esperada de ese mismo evento. En particular, las frecuencias esperadas se calcularon a partir de la integración del ajuste beta, dando como resultado un $p = 0.99$ que confirma el resultado de la prueba anterior.

Test ANOVA: Número de movimientos y aperturas

Cada juego es agrupado en una de las cinco categorías de acuerdo a la primera letra de la columna *ECO*

- H_0: La media del número de movimientos para cada grupo es el mismo.
- H_1: Al menos uno de los grupos tiene una media del número de movimientos diferente.

Inicialmente, debido a la desproporción en el número de datos por categoría, se realizó un remuestreo con reemplazo de 1500 datos por categoría. La visualización de los datos, distribución de frecuencia y boxplot, mostró que hay una variación, ligera a simple vista, de las medias de cada una de las categorías. Luego de eso se aplicó el test ANOVA sobre los datos transformados por una función logarítmica, que resultó negativo con $p = 8.53 \times 10^{-18} < \alpha$. Observando el gráfico de cajas, este resultado es consecuencia de la media más baja de C y la más alta de E; esto provee información suficiente para determinar que el número de movimientos totales en una partida de ajedrez se ve afectado por la secuencia de apertura. No obstante, se debe notar que el aumento del promedio de E puede deberse a una falta de datos. Aún así, si se tiene en cuenta que E corresponde a secuencias de apertura defensivas, se puede argumentar que el aumento de la media del número de partidas se debe a que el juego termina demorándose más en terminar su fase inicial.

Test ANOVA: número de movimientos y tipo de juego

Cada juego es agrupado en uno de los cuatro tipos de juego de acuerdo a la columna *Category*.

- H_0: La media del número de movimientos es la misma para todas las categorías.
- H_1: Al menos una categoría tiene al media del número de movimientos diferente.

Para este test se tomó nuevamente una muestra de 7500 datos con reemplazo distribuidos homogéneamente para cada una de las categorías de juego. El objetivo de este test era determinar si el número de movimientos se veía afectado por la categoría. En principio, se pensaría que las medias para cada una de las categorías es similar, porque sin importar la variación temporal del avance del juego, éste siempre avanzará de acuerdo a la ocupación de casillas centrales y la captura de piezas. El gráfico de cajas muestra que el promedio de todas las categorías está alrededor de ~ 55, resultado que se confirma con las medias de las distribuciones. El test ANOVA se realizó con los datos log-transformados, dando un valor $p = 3.49 \times 10^{-}5 < \alpha$, es decir, el test es negativo. Sin embargo, se debe notar que ese resultado es un valor que no es relevante ya que todos los promedios transformados dan lugar a ~ 4.0; el test detecta fluctuaciones por encima de la segunda cifra decimal.

Independencia estadística: Rango de elo y aperturas

Un test χ^2 es usado para determinar si las variables categóricas *EloTier* y *ECO* son estadísticamente independientes.

- H_0: Las variables son independientes.
- H_1: Las variables son dependientes.

Como en los tests previos, se utilizó una muestra con 7500 partidas con reemplazo. El test χ^2 dio como resultado un valor $p = 2.8 \times 10^{-56} < \alpha$, rechazando la hipótesis nula. Es decir, se encuentra que las variables son estadísticamente dependientes. En particular se observa el siguiente patrón para todas las secuencias de aperturas, exceptuando C: A mayor elo se juegan más partidas con una apertura en ECO. El caso de C, se explica porque la popularidad de las salidas de juego abierto disminuyen a medida que se aumenta en la escala de elo debido a la adopción de aperturas más variadas y avanzadas.

Test χ^2: Resultado y Rango de elo

Cada juego es clasificado en uno de los 4 niveles de elo de acuerdo a la media de elo entre las negras y las blancas: Low elo, Mid elo, High elo, Very high elo

- H_0: No hay una relación entre la variable *EloTier* y la variable *Result*.
- H_1: Hay una relación entre las variables *EloTier* y *Result*.

La distribución de la cantidad de jugadores por cada nivel de elo tiene un sesgo hacia la derecha, es decir, hay una desproporción en los datos entre las distintas categorías. Por lo anterior se realizó un remuestreo de 1875 datos con reemplazo distribuidos homogéneamente para cada una de las categorías del nivel de elo. Por medio de una tabla de contingencia se realizó el test de χ^2 obteniendo un valor $p = 0.01$ y una estadística de $\chi^2 = 16$ con 6 grados de libertad, rechazando la hipótesis nula. Por lo tanto las variables son estadísticamente dependientes. El porcentaje de victorias en bajo y muy alto elo son mayores en relación a las dos categorías restantes. Para elos bajos el número de errores cometidos en una partida es mucho mayor con respecto a otros elos, lo cual justifica que en estos niveles se tenga una tasa de victorias mucho mayor. Por otro lado, cuando se alcanza un muy alto elo, el conocimiento sobre aperturas, medio juego y finales reduce el número de errores cometidos en la partida por ambos jugadores, no obstante, estos a su vez generan mayores oportunidades para crear ataques efectivos y obtener una victoria. Para el Mid y Alto elo, la distribución de la cantidad de partidas ganadas y perdidas es muy similar indicando que la cantidad de errores cometidos en una partida se ve anulada por el ligero conocimiento sobre teoría del ajedrez, la cual genera oportunidades de ganar. Desde una perspectiva global, el porcentaje de victorias es mayor al de derrotas en todas las divisiones de elo. Lo anterior se debe a la ventaja existente en las piezas blancas por tener el primer movimiento a su favor. (Esto se justifica en el test A/B).

Test ANOVA: Resultado y WhiteRatingDiff

Cada categoría de *Result* tiene un conjunto de valores asociado de *WhiteRatingDiff*.

- H_0: No hay una relación entre las variables *WhiteRatingDiff* y *Result*.
- H_1: Hay una relación entre las variables *WhiteRatingDiff* y *Result*.

Dada la desproporción entre el dato *Draw* y los datos *Win* y *Lose*, se realizó un remuestreo con reemplazo de 2500 datos por cada categoría. El gráfico de violín muestra como la distribución del *WhiteRatingDiff* para la categoría *Win* se da para valores positivos y tiende a caer rápidamente cuando el *WhiteRatingDiff* se aproxima a cero. Por el contrario, sucede lo mismo en la categoría *Lose* pero se distribuye en los valores negativos y conforme *WhiteRatingDiff* se aproxima a cero esta decae rápidamente. Por último, la categoría *Draw* se distribuye alrededor del cero con un sesgo hacia los valores negativos de *WhiteRatingDiff*, lo cual tiene sentido si se considera la ventajas de las piezas blancas por comenzar las partidas.

Desde un inicio se esperaba una fuerte correlación entre las variables, dado que el *WhiteRatingDiff* contiene la información sobre una probabilidad de victoria, lo que debe de verse reflejado en la variable *Result*. El test ANOVA arrojó un valor $p = 7.5 \times 10^{-319}$ para los datos log-transformados, lo que conduce a rechazar la hipótesis nula. Este valor p tan grande solo se justifica si la relación entre las variables también es muy grande, lo cual, desde la propia definición del *WhiteRatingDiff*, se puede inferir que la variable *Result* va a estar condicionada por el *WhiteRatingDiff*.

Test A/B: ¿El color importa? (test Binomial)

En esta prueba de hipótesis se usó la prueba binomial para determinar si el color de las fichas influye en el porcentaje de victorias. Para este test se tomó un tamaño de muestra de 150k datos y se hizo un remuestreo con reemplazo.

- H_0: El porcentaje de victorias de las fichas blancas es del 50%
- H_1: El porcentaje de victorias de las fichas blancas no es del 50%

Los gráficos de distribución de la cantidad de victorias y derrotas para las fichas blancas muestran que la cantidad de victorias es ligeramente mayor a la de derrotas. Se encontró que el valor es $p = 0.004$ por lo que la hipótesis nula fue rechazada, indicando que el porcentaje de victorias de las blancas no se corresponde con un 50%. Este resultado es de esperarse debido a que en las reglas del ajedrez las piezas blancas siempre son las primeras en hacer el primer movimiento. La ventaja de tener el primer movimiento permite al jugador de las blancas elegir el tipo de apertura, lo cual condiciona el rumbo de la partida al menos en su primera parte. Desde el punto de vista de la teoría de juegos combinatorios, donde se estudian juegos secuenciales con información perfecta, se presupone que siempre hay una estrategia ganadora que lleva a la victoria a uno de los jugadores, no obstante, para casos como el ajedrez donde existe la posibilidad de tablas esto no se cumple, por lo tanto se podría esperar una tasa de 50% de victorias y 50% de derrotas si ambos jugadores, blancas y negras, hacen todos los movimientos perfectos. Sin embargo, de acuerdo a la teoría de la complejidad, el ajedrez es uno de los juegos más complejos y la posibilidad de saber cual es la mejor jugada que se puede hacer es muy difícil de calcular. Por lo anterior, la ventaja existente en las piezas blancas influye en el resultado de la partida. Imagine en un punto aleatorio durante un juego en el que de repente se le da a un jugador en particular dos movimientos seguidos, eso sería una ventaja decisiva, lo que sugiere que tener el primer movimiento debería ser realmente significativo.

Síntesis

La mayor parte de las variables que tenía la base datos eran categóricas, esto dio lugar a que se realizaran muchos tests ANOVA para observar como variaba una misma variable numérica, como el número de jugadas en la partida, con respecto a las agrupaciones de distintas variables categóricas. Así mismo, la naturaleza de la base de datos permitió realizar tests de independencia estadística entre algunas de estas variables. Teniendo en cuenta lo anterior, la mayor parte del análisis se realizó a través de distribuciones de frecuencia, esto es, conteos. Exceptuando el primer test de ajuste, todos los demás test tienen información de conteo; por ejemplo, se encontró que para dos categorías de ECO, C y E, la media del número de jugadas difiere significativamente del resto. Así mismo, la tabla de contingencia elo-eco muestra cómo varía el conteo de las categorías de elo de acuerdo a las categorías de ECO (ver cuaderno de Jupyter). Quizás el resultado más destacado corresponde al obtenido por el test binomial, se encontró que los jugadores con piezas blancas ganan más a menudo; esto es un hecho que no encuentra una explicación en la distribución de datos, y resulta es de las mismas reglas del juego: el jugador que inicia tiene una ventaja.

Uno de los próximos pasos para un análisis de esta base de datos es realizar test de hipótesis en compañía de una persona experta en la teoría del ajedrez, que pueda ayudar a justificar las relaciones existentes entre las variables desde la teoría y su experiencia. Por otro lado, es recomendable realizar otros tipos de hipótesis para analizar relaciones entre las variables como victorias y aperturas o elo y número de movimientos. Un próximo análisis de dato sobre las partidas de ajedrez en lichess, es usar varias bases de datos de diferentes meses y años, no solo se podrían realizar nuevas hipótesis de relaciones entre las distintas bases si no que se aumentaría la población de las distintas variables y se podrían hacer mejores remuestreos para así, los

test que tuvieron algunos problemas con la cantidad de datos como lo fueron los ANOVA, puedan verse solucionados.

Ahora, Se debe mencionar que el conjunto de datos tenía todas las características usuales que se pueden obtener de una partida de ajedrez, además de estar muy bien organizado. Sin embargo, respecto a la realización de pruebas de hipótesis algunas de estas no resultan ser particularmente innovadoras, en tanto que la cantidad de variables numéricas relevantes en relación al número de variables categóricas, no permite flexibilizar las opciones que se tienen para realizar tests. Ahora, para realizar ANOVAS siempre se tiene que recurrir a un remuestreo con los datos que se tienen debido a la desproporción en cuentas entre algunas categorías. Esto resulta en que las variaciones de medias que se detectaron significativas pueden ser sólo el resultado de un submuestreo de categorías. Por otro lado, debe mencionarse que el número de partidas por remuestro se tomó de forma arbitraria, aún cuando se buscaba que hubiera un balance entre sensibilidad de los tests y el riesgo de submuestreo: se requiere revisar si había algún parámetro que permitiera calcular esto con rigor. En últimas, se requiere un conjunto de datos del mismo tamaño para cada categoría. Por último, el tipo de información que provee la base de datos no es apta para realizar machine learning, aun cuando se podrían realizar modelos de regresión para predecir, por ejemplo, si gana blanco de acuerdo a todos los parámetros independientes que se tienen, en última instancia los únicos parámetros relevantes son las rating calculados y asignados antes de cada partida.