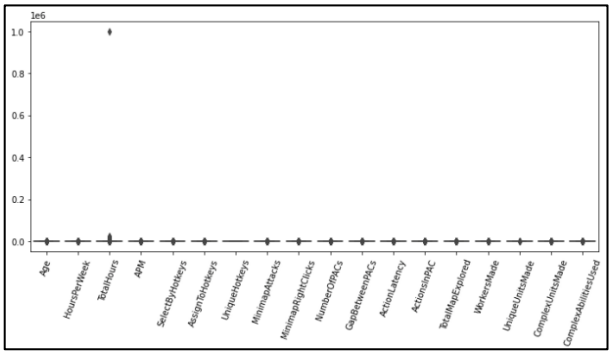
María Camila Terán — 201822000
Juan Diego Cardona — 201819447
Nicolás Ortega — 201814515

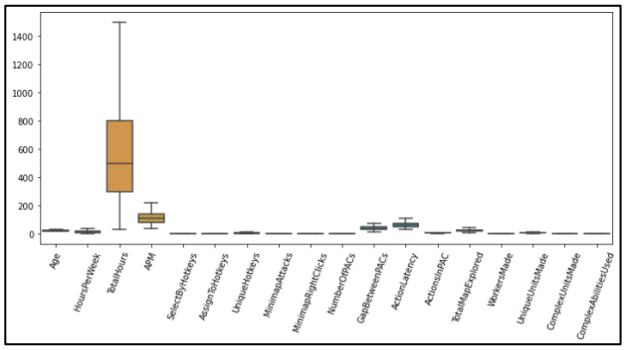
Inteligencia de Negocios
Laboratorio 3

Limpieza y pre-procesamiento – Nicolás Ortega

En primer lugar, se realizó una revisión y análisis de los datos suministrados en el archivo .csv. Preliminarmente, encontramos que existían 19 columnas, de las cuales la primera correspondía a la variable objetivo (LeagueIndex) y el resto a las variables explicativas que se utilizarían para construir el modelo. El número total de registros en los datos era de 3238. En esta ocasión no existían registros con valores faltantes o nulos. Sin embargo, en este perfilamiento sí se vio que existían valores inconsistentes en casi todas las columnas, pues aunque todas debían contener valores numéricos, se estaban leyendo como objetos o cadenas de texto. Luego de una revisión, se encontró que algunos registros tenían valores escritos como "NA_VALUES", esto era lo que estaba generando el error. Adicionalmente, se notó que en la columna de la variable objetivo también existían valores incoherentes, ya que en varios registros estaba reportado el valor de 20, lo cual no podía suceder de acuerdo con el diccionario de datos. Para solucionar estos problemas se eliminaron los registros que contuvieran los valores mencionados.

Una vez se solucionaron estos problemas, se procedió a revisar los valores de los datos y a manejar los valores atípicos. Luego de la limpieza, se revisó la distribución de los datos de cada columna y se encontró que la mayoría contenía una cantidad considerable de valores atípicos (en especial la columna *TotalHours*). Dado que estamos en un contexto de regresión lineal, era necesario lidiar con estos valores para que no generaran una distorsión en el modelo. Para solucionar este problema se utilizó el método de *Winsorizing*, que consiste en poner un límite mínimo y máximo para los valores (siempre iguales a o entre el primer y tercer cuartil de la distribución original) y pasar los atípicos a estos valores. Luego de aplicar el método las columnas no contaban valores atípicos:

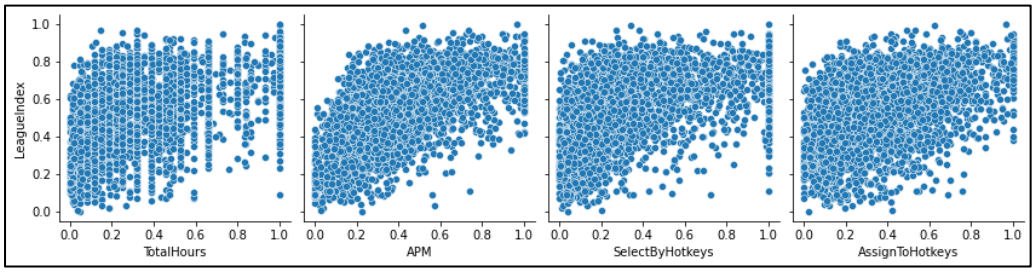


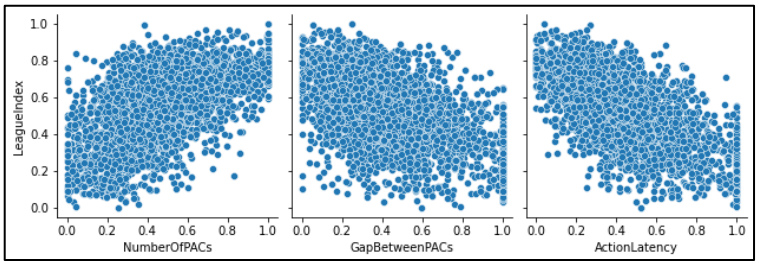


Como se puede observar, las variables explicativas manejan escalas muy diferentes en sus valores. Algunas son muy pequeñas y otras mucho más grandes. Por esto, y para finalizar la fase de pre-procesamiento de datos, se realizó una normalización para que todas tuvieran valores entre 0 y 1.

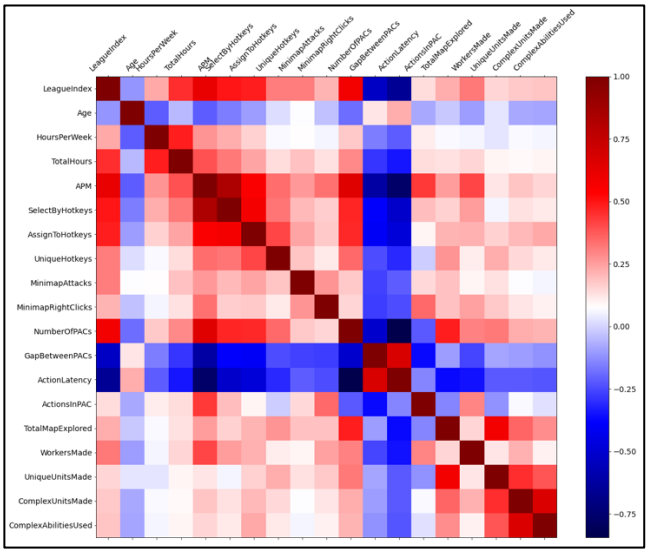
Identificación de variables – Nicolás Ortega

Una vez realizada la limpieza de los datos, se procedió con la selección de las variables explicativas que se utilizarían para construir el modelo. Primeramente, decidimos visualizar gráficamente cómo se relacionaban las variables dadas en los datos con la variable objetivo. Para esto, se realizaron gráficos de dispersión de cada una y se identificaron algunos candidatos que mostraban una posible relación existente con la variable objetivo:





En los gráficos podemos ver que las variables *'TotalHours','APM','SelectByHotkeys','AssignToHotkeys', 'NumberOfPACs'* parecen estar directamente relacionadas con la variable objetivo, mientras que las variables *'GapBetweenPACs', 'ActionLatency'* parecen estarlo inversamente. Para verificar un poco esta hipótesis, se observó una matriz de correlaciones:

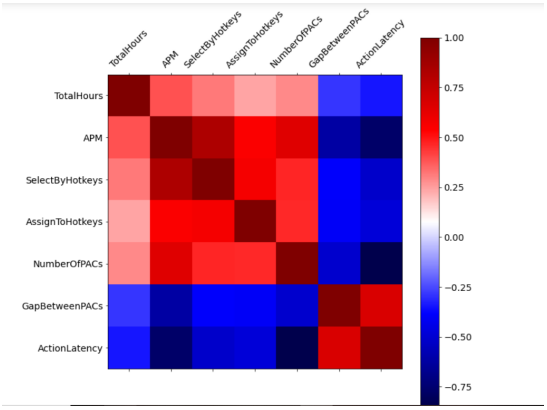


Al ver esta matriz verificamos que los candidatos seleccionados sí tienen una correlación mayor que el resto de las variables y decidimos entonces utilizarlos para implementar el modelo de regresión.

Verificación de supuestos – María Camila Terán

1. Colinealidad

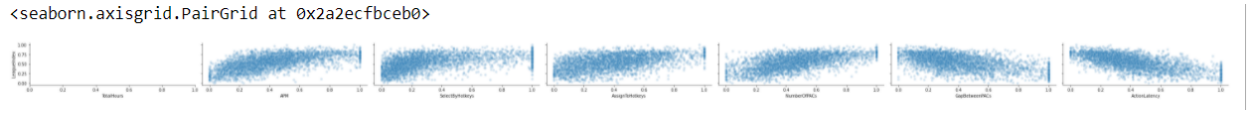
Para verificar el supuesto de la colinealidad, se realizó la matriz de correlación que dio como resultado lo siguiente:



Como se puede observar, hay una baja colinealidad entre las columnas utilizadas por lo que las variables utilizadas son pertinentes.

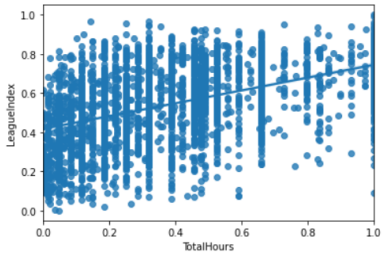
2. Linealidad

Se espera que haya una relación lineal entre las variables utilizadas. Esto se demuestra en la siguiente representación gráfica:



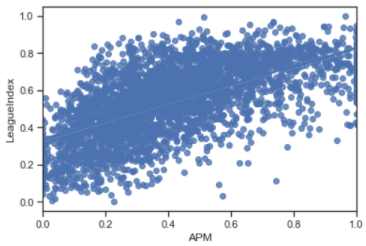
De igual forma, para estar más seguros, se halló el coeficiente de correlación Pearson con el fin de asegurar la linealidad. Este va desde -1 a 1 pero se espera que su valor sea positivo.

TotalHours:



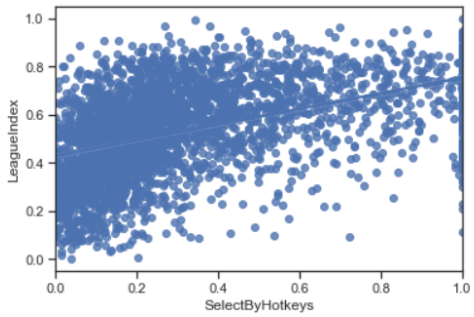
Coeficiente Pearson: 0.45

APM:



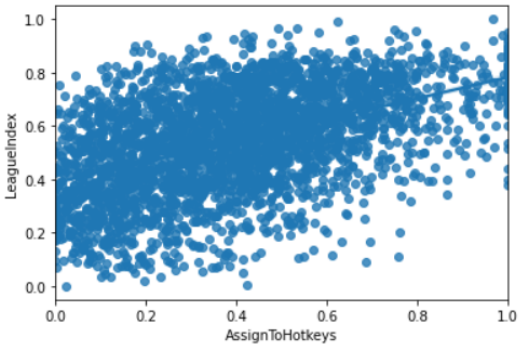
Coeficiente Pearson: 0.62

Selected by hot keys:



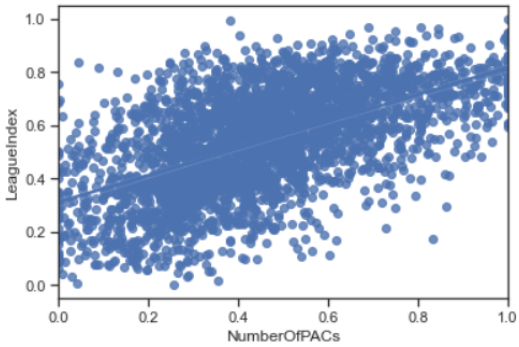
Coeficiente Pearson: 0.50

Assigned to hot keys:



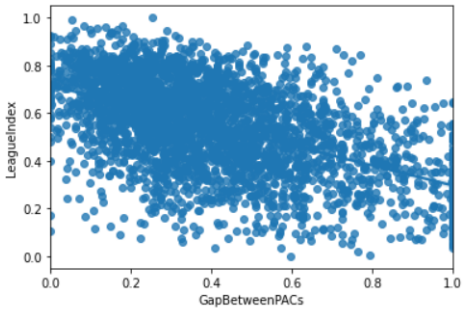
Coeficiente Pearson: 0.48

Número of PACs:



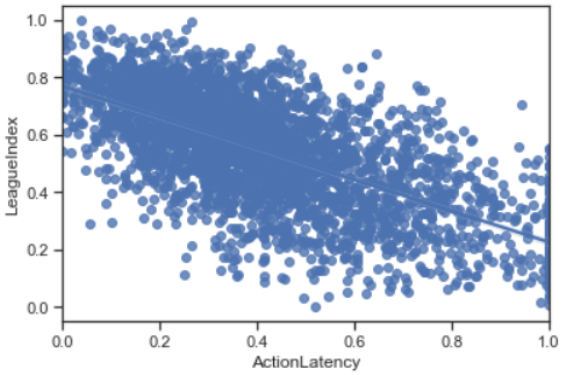
Coeficiente Pearson: 0.58

Gap between PACS:



Coeficiente Pearson: -0.528

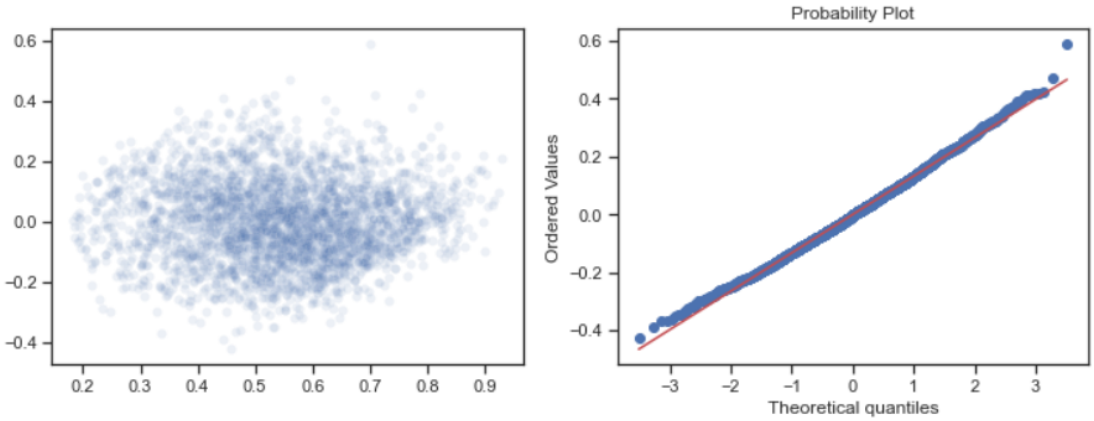
Action Latency:



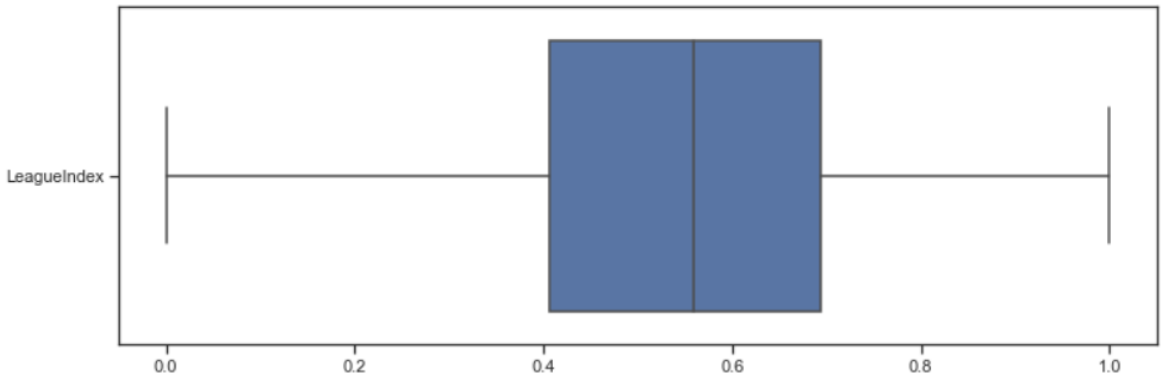
Coeficiente Pearson: -0.654

3. Errores

Como se busca minimizar el error, se gráfica lo siguiente:



Efectivamente las probabilidades son una línea recta que es lo esperado. Además, la media concuerda y no hay datos atípicos que puedan distorsionar el modelo. Esto se ve a continuación:



Análisis de resultados, interpretación de coeficientes – Juan Diego Cardona

Se obtuvieron los siguientes coeficientes:

```
array([ 0.15929791,  0.01550315,  0.06167175,  0.11078287,  0.07487103,
       -0.09596032, -0.24479904])
```

De acuerdo con los resultados de la regresión se obtuvieron las siguientes resultados del error cuadrático medio y el coeficiente de determinación para medir su confiabilidad:

$$R^2 = 0.53$$

$$\text{RMSE} = 0.137$$

De esto se puede concluir que en efecto hubo mejoras respecto al modelo entregado por la empresa, no obstante, se espera aumentar el valor. Sin embargo, el error si se minimizó como era lo esperado lo cual indica que se obtuvo un buen modelo final.

Las variables mas influyentes fueron 'TotalHours','APM','SelectByHotkeys','AssignToHotkeys', 'NumberOfPACs' pues son las que están mas relacionadas con la variable objetivo

Mientras que las variables 'GapBetweenPACs', 'ActionLatency' tienen una relación inversa entre ellas.

Se encontró una relación lineal con las variables APM, SelectByHotkeys, AssignToHotkeys. Y una relación inversamente lineal con apBetweenPACs ActionLatency.

Se recomienda instalar el modelo de estimación en producción.