Brahayan Stiven Gil Henao
Jhon Alexander Bedoya Carvajal
Sebastián Salamanca Mendez

Informe Resultados Caso de Estudio - Aprendizaje Supervisado

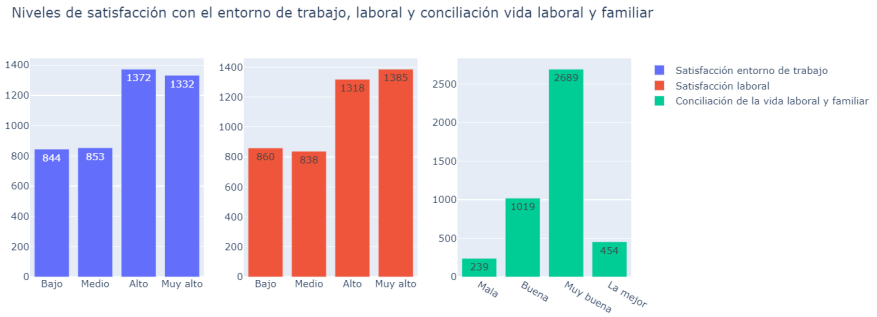
CONTEXTO

Limpieza y Transformación

Para implementar la solución de la problemática se contaba con 4 bases de datos las cuales fueron unidas en una sola base, además, se realiza la transformación de los nombres de las variables para facilitar su manejo. Las variables categóricas estaban codificadas numéricamente, se les hace la asignación de sus categorías correspondientes a cada una de ellas para facilitar el entendimiento del conjunto de datos, por otra parte, algunas categorías presentaron problemas en su escritura las cuales fueron corregidas. Se hace el tratamiento de valores faltantes, las variables categóricas se rellenaron con la moda y para las variables numéricas se fueron rellenadas con la mediana, para finalizar con la limpieza de los datos, se eliminaron 3 columnas constantes y se exporta el dataset limpio.

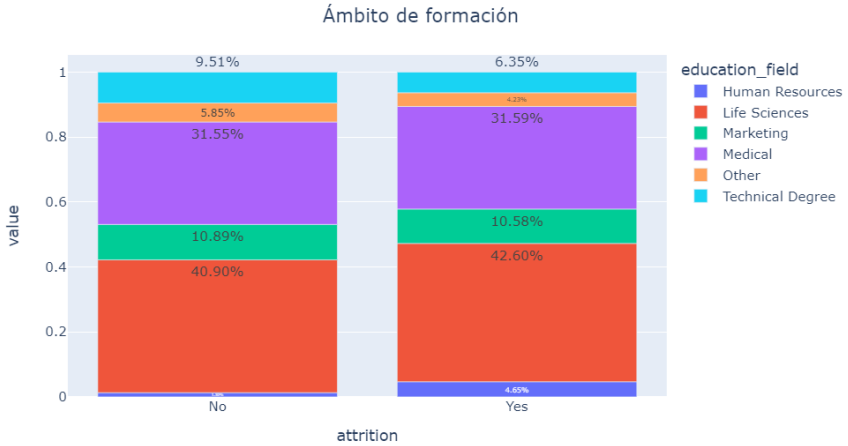
Análisis Exploratorio de Datos (EDA)

Se continua con el análisis univariado de las variables categóricas, para este análisis se realizan gráficos de barras, boxplot y tablas de frecuencia para cada variable individualmente donde se evidencia que la variable objetivo tiene una gran predominancia del no abandono del puesto de trabajo, lo que se esperaría debido a que las personas normalmente no cambian su puesto de trabajo de un año para otro. También se evidenció que las puntuaciones para las variables “Satisfacción con el entorno de trabajo” y “Satisfacción laboral” tienen una distribución casi igual, es de esperarse que un empleados satisfecho con su entorno de trabajo tenga también una alta satisfacción laboral, pero en ambos aspectos hay un porcentaje alto de empleados que califican en bajo y medio, por lo que la satisfacción de los empleados no está en un buen punto. Asimismo, todos los empleados tuvieron un rendimiento “Sobresaliente” y “Excelente” en el último año, siendo calificados la gran mayoría,con un rendimiento “Excelente”, lo que sugiere que la empresa está conforme con el rendimiento que dieron los empleados en el último año. Por otro lado se tendría que entrar a analizar cómo están siendo medidos ya que no es lo más normal que todos tengan un rendimiento sobresaliente.



Para el análisis individual de las variables numéricas se utilizaron boxplot, histograma y estadísticos descriptivos como la media, cuartiles, mínimos y máximos, se observa que las variables relacionadas con los años presentan un gran número de valores atípicos en el extremo superior, lo que significa que hay personas que llevan muchos años en la empresa, pero la mayoría de observaciones se acumulan en valores bajos, esto se debe a la problema de rotación actual de la empresa.

En el análisis bivariado de las variables categóricas se realizaron tablas cruzadas y diagramas de barras de las variables frente a la variable objetivo, donde se logra observar que el pertenecer al departamento de recursos humanos aumenta las probabilidades de renunciar a la empresa así como, viajar frecuentemente por motivos de trabajo, que su ámbito de formación se los recursos humanos, estar en estado civil soltero,tener bajos niveles de satisfacción laboral, con el entorno de trabajo y conciliación entre la vida familiar y laboral.



Continuando con la exploración del análisis bivariado de las variables numéricas se observa que las personas que renunciaron tienen un promedio de edad menor que las que no lo hicieron y también dedicaban más horas de trabajo diarias, además habían pasado anteriormente por una mayor cantidad de empresas y llevaban menor cantidad de tiempo en su etapa laboral dentro y fuera de la compañía.

Análisis de Correlaciones

En el análisis de correlaciones numéricas se encuentra una alta relación entre las siguientes variables:

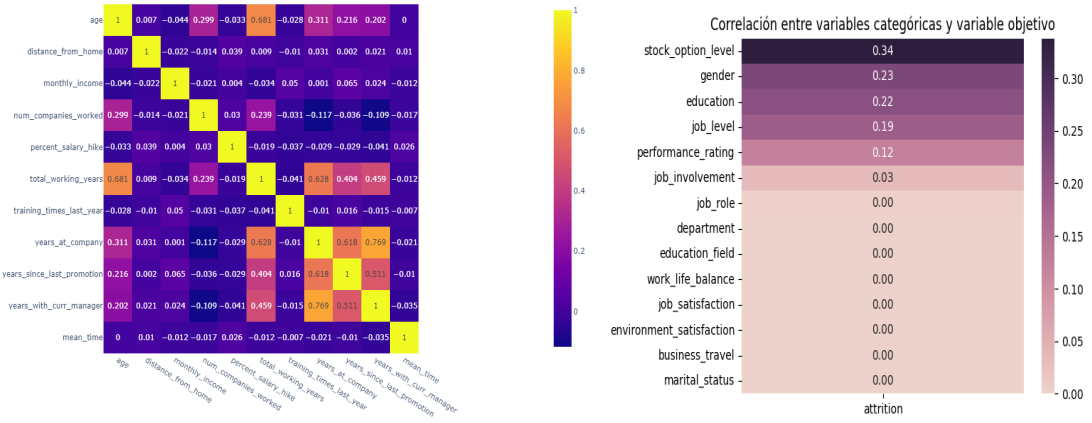
- Entre edad y total de años trabajando, lo cual tiene sentido que los empleados de más edad sean los que lleven más años trabajando en la empresa.
- Entre total de años trabajando y total de años en la compañía. Los empleados que llevan más años como empleados tienden a ser los que llevan más años en la compañía.
- Entre años en la compañía y años desde la última promoción. Entre más tiempo tenga el empleado en la compañía es más probable que haya pasado mucho tiempo desde su último ascenso.
- Entre años en la compañía y años al mando del jefe actual.

Todas las variables relacionadas con el tiempo de estancia del empleado en la empresa tienen una correlación alta entre ellas. 'Años en la compañía' y 'Años con el jefe actual' son muy parecidas estadísticamente hablando y plantea la idea de eliminar una de estas, se propone usar otros métodos de selección de variables para analizar los resultados.

En el análisis de correlación de las variables numéricas frente a la objetivo se encontró que las variables “Distancia desde casa”, “Porcentaje de aumento salarial” y “Veces que fue entrenado el empleado en el último año” no influyen en el abandono del empleado.

Los empleados que sí abandonaron tenían ingresos mensuales promedio inferiores a los que no abandonaron, a pesar de tener el mismo rol y tener mayor promedio de tiempo de trabajo al día.

Por otro lado en el análisis de correlación de las variables categóricas se encontró que no hay evidencias de que las variables Educación, Género, Nivel del puesto, Nivel de opciones sobre acciones del empleado y Valoración del rendimiento en el último estén relacionadas con la variable objetivo (el empleado abandonó el año anterior) por lo que no aportaran mucho a predecirla.



Preparación de datos

Varias características de naturaleza numérica presentaron valores muy extremos. Para tratar con estos valores se optó por reemplazar sólo los valores que se salieran de 2 rangos intercuartílicos e imputar estos con la mediana por grupo, donde los grupos eran si el

empleado abandona o no; de esta manera se alteró lo menos posible la naturaleza de los datos.

Además, se realizaron otro análisis gráficos relevantes para determinar si los valores extremos presentaban alguna dependencia con otra característica, encontrando una respuesta negativa a esto. Por último, las variables de naturaleza categórica se dummizaron.

Selección de características

En la selección de características se realiza una selección manual incluyendo de las variables relacionadas con los años solo la variable “Años en la empresa” ya que estas estaban altamente relacionadas entre sí y excluyendo las variables categóricas “Educación” , “Género”, “Nivel del puesto”, “Nivel de opciones sobre acciones del empleado” y “Valoración del rendimiento en el último y las variables numéricas “Distancia desde casa”, “Porcentaje de aumento salarial” y “Veces que fue entrenado el empleado en el último año” por que no estaban relacionadas con la variable objetivo, se utilizaron 4 técnicas de selección de variables para crear 9 subconjuntos de datos, 2 por cada técnica y el conjunto seleccionado manualmente, 4 conjuntos con 30 características,1 con 31, 1 con 27, otro con 16 y uno más con 11 y el manual con 45 características, todos los conjuntos de datos se probaron con dos modelos cada uno para analizar las métricas dando como ganador a los conjuntos seleccionados manualmente y a través del método anova

Selección de modelos

Se procedió con los dos conjuntos seleccionados anteriormente más el conjunto con todas las características para probar 3 modelos y optimizar los hiperparametros de cada uno de estos, finalizando con la elección del mejor modelo que fue el desarrollado con el conjunto de datos seleccionados con el método anova y con el modelo “Xtreme Gradient Boosting Classifier” ya que tenía unas métricas muy similares a las del mejor modelo que fue logrado con la selección manual y el “Xtreme Gradient Boosting Classifier” pero contando con 15 características menos que este, lo que lo hacia un modelo mucho más simple.

Conclusiones y recomendaciones

- Al tener modelos perfectos en entrenamiento, ambos modelos presentan sobreajuste, pero en este caso no tiene efectos negativos, dado que en las pruebas los modelos siguen arrojando muy buenos resultados en las métricas de desempeño.

- Se selecciona el modelo “Xtreme Gradient Boosting” con la selección obtenida con el método ANOVA. Aunque el modelo “Xtreme Gradient Boosting” con la selección manual presenta métricas un poco mejores, la diferencia es muy poca teniendo en cuenta la complejidad de ambos modelos, el modelo con la selección manual fue

ajustado con 45 características, mientras que el modelo seleccionado fue ajustado con 30 características.

- La característica más importante para el modelo seleccionado (“Xtreme Gradient Boosting Classifier” con ANOVA) es la educación en el campo de Recursos Humanos. Para combatir el problema de abandono de empleados, poner especial atención en los que su ámbito de educación es en Recursos Humanos.

- El modelo indica que los empleados calificados con una baja implicación en el trabajo y los empleados que dieron una calificación de "Mala" al balance vida-trabajo también tienen mucho peso cuando se quiere predecir si un empleado va a abandonar.

- Hay que concentrarse en el área de recursos humanos ya que se ve que tanto el campo de educación como el departamento está afectando la decisión de abandonar, hay que buscar formas de incentivar a los trabajadores o de nivelar sus condiciones de trabajo con las de otros departamentos y campos de saber.

- Se evidencia que las personas que tomaban la decisión de abandonar la empresa trabajaban más horas en promedio diarias y obtienen menos ingresos,también que eran personas mas jóvenes y que llevaban menos tiempo en la empresa, se podría pensar que por ello ganaban menos o trabajaban más horas, pero estas variables no están relacionadas entonces hay que hacer una investigación exhaustiva del porque estas personas estaban trabajando más y ganando menos.