INFORME CASO DE ESTUDIO SOBRE RRHH

DIEGO ALEJANDRO SAAVEDRA VALDIVIESO

NILSON SUAREZ HERNANDEZ

DAYAN EDUARDO MARÍN QUINTERO

APLICACIONES DE LA ANALÍTICA

PROFESOR:

JUAN CAMILO ESPAÑA LOPERA

INGENIERÍA INDUSTRIAL

FACULTAD DE INGENIERÍA

UNIVERSIDAD DE ANTIOQUIA

2023

a. Comprensión del problema de negocio y traducción a problema analítico

Problema de negocio: El problema de negocio que enfrenta esta empresa se relaciona con la alta tasa de retiros de empleados, que está en torno al 15% anual. Esta situación tiene varias implicaciones negativas y costosas para la empresa.

Impactos Negativos Actuales: La alta rotación de empleados conlleva costos financieros considerables, retrasos en proyectos, una mayor carga de trabajo en la selección de personal, sobrecarga a los empleados restantes y resulta en la pérdida de conocimiento y experiencia valiosa

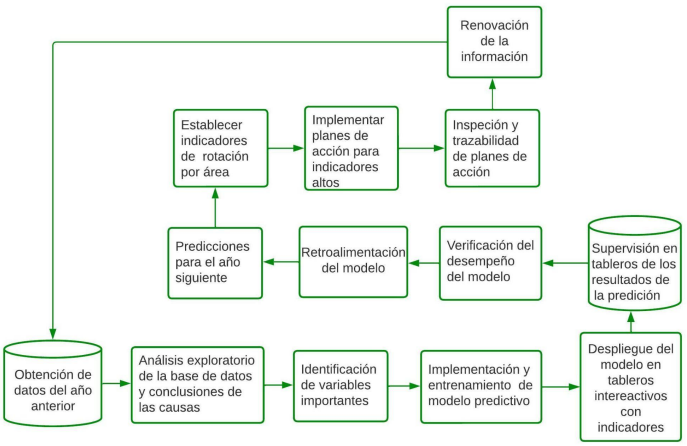
Objetivo Principal: Reducir la tasa de retiros de empleados en la empresa, actualmente se encuentra en un 15% anual, con el fin de mitigar los costos asociados y minimizar las implicaciones negativas en el funcionamiento de la organización.

Beneficios Esperados: Reducir la rotación de empleados conlleva ahorros financieros, estabilidad laboral, evita retrasos en proyectos y mejora la satisfacción de clientes y empleados al conservar el conocimiento y la experiencia.

Métricas de Éxito: La reducción de la tasa de retiros anual, respaldada por la precisión de las predicciones de nuestro modelo, no solo tiene un impacto positivo en los costos operativos y la eficiencia laboral, sino que también garantiza el mantenimiento de plazos y la calidad en proyectos. Al retener a nuestros empleados clave, estamos fortaleciendo la retención de conocimiento y experiencia dentro de la organización, lo que contribuye significativamente al crecimiento y éxito continuo de la empresa.

Problema analítico: Analizar las variables generales que influyen en la deserción de los empleados y tomar acciones correctivas sobre estas variables con el fin de lograr condiciones óptimas en cuanto a la satisfacción de los empleados y mitigar las consecuencias que esto trae sobre la empresa. Así mismo realizar predicciones para ver si el empleado abandonará o no la empresa con el fin de tomar acciones preventivas sobre un empleado puntual que muestre índices de inconformidad. En este caso es de suma importancia realizar dicho proceso con inteligencia artificial dado que se ahorraría mucho tiempo al procesar los más de 4000 registros que al realizarlos de forma manual

b. Diseño de solución propuesto.



c. Limpieza y transformación.

Se renombran algunas columnas, se cambia el formato de aquellas que tienen fechas y se eliminan las variables "Over18", "StandardHours", "EmployeeCount", ya que no generan un valor agregado al análisis. En la columna “resignationReason” (motivo de renuncia), los datos nulos son reemplazados por el valor “Fired” porque estos corresponden a las personas que fueron despedidas. Algunos son eliminados por tener una cantidad baja, como es el caso de "NumCompaniesWorked" con 19 y "TotalWorkingYears" con 9 datos. Y para los nulos de EnvironmentSatisfaction, JobSatisfaction y WorkLifeBalance se reemplaza por la moda de ese empleado en el departamento que se encuentren

d. Análisis exploratorio.

Se carga el dataframe guardado en sql, luego se utilizan querys para obtener a simple vista un panorama de la empresa respecto a las posibles causas de deserción, se obtienen los promedios de ambiente laboral por departamento: Human Resources, Research & Development y Sales, donde se pudo evidenciar que los los valores eran de 2.82, 2.71 y 2.72 respectivamente, podemos inferir de una manera anticipada qué el ambiente laboral en las áreas de la empresa no es muy bien calificado por los empleados. Luego de lo mencionado, se realiza una matriz de correlación con la cuál se quiere identificar aquellas variables que pueden tener algún tipo de relación entre sí, adicionalmente se quiere conocer cuales de estas influyen en mayor magnitud en que un empleado abandone o no la empresa y se identifican las siguientes:

- El tiempo medio que pasa en el trabajo (entre más tiempo pase, más probable es que abandone la empresa)
- Total de años que a trabajado en toda su vida (entre más años, menos probable es que abandone la empresa)
- La edad (entre más joven, más probable es que abandone la empresa)
- Años con el actual jefe (entre más años, menos probable es que abandone la empresa)
- Satisfacción con el trabajo general (entre más satisfecho, menos probable es que abandone la empresa)
- Satisfacción con el entorno laboral (entre más satisfecho, menos probable es que abandone la empresa)

Se realizaron los gráficos de las variables numéricas con la variable respuesta, mediante el uso de boxplots, algo a destacar es que efectivamente hay más empleados que en promedio tienen un tiempo promedio mayor en la empresa y estos son los que en su mayoría abandonan, estos datos tienen mucho sentido en el contexto real. Además los jóvenes son los que más tienden a renunciar, posiblemente por eso las variables TotalWorkingYears, Age, YearsAtCompany, Age y YearsWithCurrManager tienen una correlación fuerte, ya que los jóvenes tienen menos años trabajados en su vida, menos años en la empresa, menos años con el jefe actual y menos edad. Finalmente respecto a las variables categóricas podemos tener las siguientes hipótesis, el perfil de una persona que abandona rara vez viaja, son de Investigación y Desarrollo, estudian ciencias de la vida y medicina, son hombres y están solteros. Cabe aclarar que estas afirmaciones pueden estar sesgadas pero es importante tenerlas en cuenta desde un principio.

e. Selección de algoritmos y técnicas de modelado.

Utilizaremos la división 70-30 para tener un modelo más equilibrado que tenga suficientes datos para entrenar y para testear y así poder dar un veredicto más confiable

Para este caso particular, se eligen 5 modelos: LogisticRegression, DecisionTreeClassifier, RandomForestClassifier, GradientBoostingClassifier y XGBClassifier, con validación cruzada de 5 folds.

Vemos que la clase target está desbalanceada, porque el dataset tiene más registros de gente que no abandona que de gente que si abandona, por lo que en los modelos se usará el parámetro para balancear las clases.

f. Selección de variables.

Se eliminan las variables retirementDate, retirementType y retirementReason, ya que estas solo tienen registros de abandono y pueden generar un sesgo. Con esto, la base de datos queda con 25 variables explicativas y 1 variable objetivo.

g. Comparación y selección de técnicas.

La métrica que usaremos será el recall, teniendo en cuenta que por las características del problema buscamos reducir al máximo que el algoritmo predice que alguien no se va a ir cuando realmente si se va a ir (falsos negativos) ya que esto le genera un costo muy elevado a la empresa y por otro lado, que el algoritmo predice que alguien se va a ir cuando realmente no se va a ir (falsos positivos) no genera un costo tan elevado como el anterior.

De los 5 modelos propuestos y teniendo en cuenta la métrica seleccionada, se pueden observar los siguientes valores.

Modelo	**Recall mean train**	**Recall mean test**
LogisticRegression	0.778 +/- 0.012	0.741 +/- 0.029
DecisionTreeClassifier	1.000 +/- 0.000	0.867 +/- 0.039
RandomForestClassifier	1.000 +/- 0.000	0.843 +/- 0.032
GradientBoostingClassifier	0.636 +/- 0.014	0.485 +/- 0.021
XGBClassifier	1.000 +/- 0.000	0.882 +/- 0.038

Fuente. Elaboración propia

De los resultados, se identifica que el modelo que mejor se adapta a las necesidades del problema es el XGBClassifier, ya que tiene un recall alto en train (1) y en test (0.882), además, tiene un nivel medio de sobreajuste.

h. Afinamiento de hiper-parámetros.

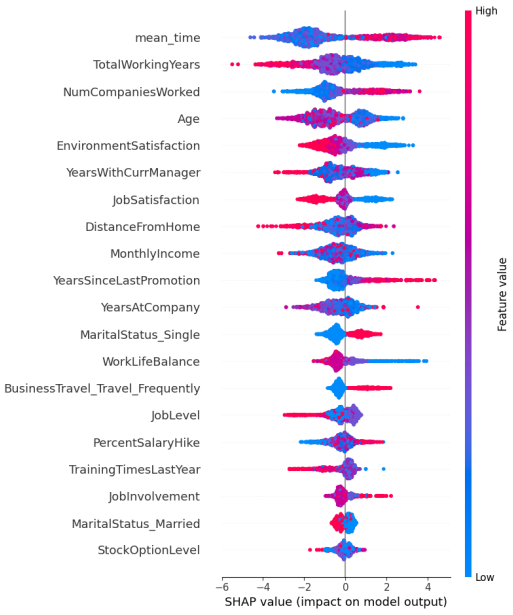
Con la finalidad de mejorar el desempeño del modelo se analizan aquellos parámetros que pueden ser significativos y que con su afinamiento se obtenga un mejor resultado, dicho esto, se hace una búsqueda por medio de GridSearchCV y validación cruzada con 5 folds. Afinando diferentes parámetros max_depth, gamma, min_child_weight, learning_rate,

subsample, reg_lambda, reg_alpha en una primera instancia se analizan **2160** modelos y el mejor arroja un **recall de 0.905**, para la segunda corrida con **1200** modelos se obtiene un **recall de 0.909** con el mejor modelo.

i. Evaluación y análisis del modelo.

Al tomar el modelo con los parámetros afinados se realizan las predicciones y se obtiene un **recall de 0.963** en un conjunto de 1315 observaciones, al ver la matriz de confusión vemos que a 7 empleados predijo que se iban a quedar cuando realmente se iban y 13 empleados indicó que se iban cuando realmente se quedaban en la empresa. Hay una diferencia y es porque el recall busca reducir esos falsos negativos Adicionalmente, para ver qué variables son las que más influyen en este resultado se utiliza la librería SHAP, que nos dice que características son más importantes para el modelo en general y también el impacto que tienen en la predicción final (si es fuerte o débil)

Características más importantes del modelo:



Nota: Nos referimos a "impacto", cuando la variable arrastra fuertemente la predicción al abandono del empleado (ya que en algunos casos la variable no tuvo tantos registros para quedar en primeros lugares pero los pocos que tuvo impactaron fuertemente, es por eso que no siempre están en los primeros lugares, pero si la cola de color rojo se extiende mostrando su impacto. (si no mencionamos impacto, es porque es un impacto bajo o medio)

En resumen podemos decir que: tiempos altos en la empresa (impacto muy fuerte), pocos años trabajados en su vida (impacto muy fuerte), que haya trabajado en muchas compañías (impacto fuerte), que sea joven, que tenga una satisfacción y un ambiente laboral bajo, que tenga pocos años con su jefe actual, que viva cerca de la empresa (impacto fuerte), que hayan pasado muchos años desde el último ascenso (impacto muy fuerte), que tenga pocos años en la compañía, que sean solteros, que no califique bien su relación trabajo-vida, que viajen con frecuencia, que tenga pocas funciones en la empresa, que tengan un porcentaje de aumento salarial alto, que haya tenido pocas capacitaciones el último año (impacto fuerte) son el perfil de un empleado que abandona la empresa.

Análisis: Tiempos altos en las empresas generan fatiga laboral y conduce al abandono de la empresa, que sean jóvenes se relaciona con otras variables, porque si son jóvenes, llevan pocos años trabajados en su vida, tienen pocos años con su jefe, pocos años en la compañía y posiblemente sean solteros, su explicación puede tener sentido ya que si son jóvenes están iniciando su vida laboral y están buscando experiencia laboral antes de quedarse fijos en una compañía.

Que haya trabajado en muchas compañías, se explica porque si es un empleado que ha trabajado en muchas compañías es un indicio que no tiene una estabilidad laboral por lo tanto

puede tender a renunciar en la empresa actual. El ambiente y satisfacción laboral baja, es muy común que cuando los empleados no están satisfechos tienden a renunciar.

Posiblemente los empleados no están sintiendo valor en dicha empresa, al no ser ascendidos, asignarles pocas funciones y que obtienen pocas capacitaciones al año, por eso pueden estar buscando otros trabajos

j. Despliegue del modelo

Se tendrá un proceso completamente automatizado en el que se tomará la base de datos de empleados del mes anterior, vigilada por RH y será enviada al modelo alojado en el sitio web corporativo, el cuál solo tendrá acceso un área nueva llamada "**RH Analytics**", la cuál consta de un equipo de 5 personas que está conformado por científicos de datos y personal de recursos humanos para juntos crear estrategias y hacer capacitaciones a los empleados. Dicho tablero mostrará indicadores de deserción, además, como se va a ir actualizando mensualmente, se verá el efecto de las estrategias implementadas, logrando no solo un impacto en la deserción sino también un impacto en la empresa de manera global

Monitoreo del modelo: Será re-entrenado automáticamente cada semestre y mediante un tablero digital alojado en el mismo sitio web corporativo, podrá tener acceso el área de " RH Analytics", pero este tablero entregará información respecto al modelo: curvas de aprendizaje, matrices de confusión y métricas del modelo, para que puedan comprobar y decidir si a ese modelo entrenado se le deben hacer ajustes o si está apto para el despliegue

k. Conclusiones

- Se tiene un modelo con un 96.3 % de confiabilidad con el cuál se crearán estrategias de la mano de recursos humanos para impactar directamente la deserción de empleados
- Al implementar dichas estrategias estamos seguros que se logrará reducir la deserción laboral
- Es de suma importancia re entrenar el modelo semestralmente con el fin de ir monitoreando su desempeño

l. Recomendaciones

- Es importante recordar que las jornadas laborales de los empleados deben ser estables y no exceder un nivel elevado de trabajo
- Crear estrategias para que los jóvenes perduren en la empresa, utilizando incentivos como bonos para postgrado, créditos para vivienda, becas, etc.
- Evitar contratar personal que ha trabajado en muchas empresas o preguntar cuál fue la causa de su renuncia en sus puestos de trabajo
- Implementar en la empresa encuestas más detalladas y con mayor frecuencia de satisfacción laboral y ambiente laboral para conocer por qué estas están en promedio tan bajas y así evitar la deserción laboral
- Brindar valor a los empleados, asignarle funciones relacionadas con su puesto de trabajo, más capacitaciones al año, esto genera que los empleados se sientan satisfechos
- Revisar que empleados son posibles candidatos a un ascenso y crear una ceremonia con los empleados para motivarlos

Repositorio en Github