ANÁLITICA EN RECURSOS HUMANOS

ESTRATEGIAS PARA LA TOMA DE ACCIONES CON LA FINALIDAD DE REDUCIR EL PORCENTAJE DE RETIROS EN LA COMPAÑÍA

APLICACIONES DE LA ANÁLITICA DE DATOS

Juan José Acevedo Dávila

Juan Sebastián Zuluaga Montenegro

Fabian Gómez Loaiza

Septiembre de 2023

Introducción

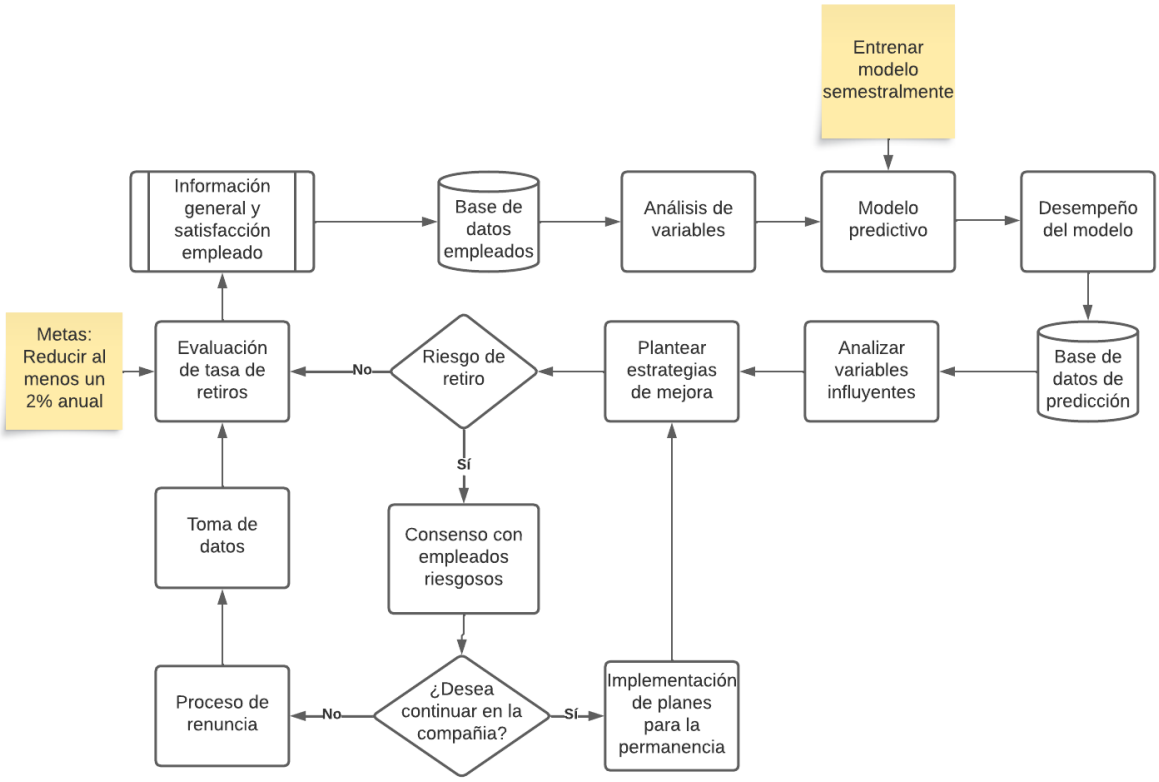
Descripción del problema

Una empresa con alrededor de 4000 empleados enfrenta una alta tasa de retiros anuales del 15%. Esta situación genera costos significativos, incluyendo gastos de contratación, capacitación y atrasos en proyectos. Además, tiene implicaciones negativas como el aumento del trabajo en el área de selección, la carga adicional para los empleados que permanecen y la pérdida de conocimiento y experiencia. La dirección de la empresa busca soluciones para abordar este desafío.

Diseño de solución propuesto

Objetivo

Desarrollar estrategias efectivas para reducir la tasa de retiros de empleados en la empresa y, en última instancia, mitigar los costos financieros y operativos asociados con la alta rotación de personal. Esto implica mejorar la satisfacción laboral, retener el conocimiento y la experiencia de los empleados, y garantizar la eficiencia en los procesos y proyectos.



Metodología

Fuentes de datos

Para llevar a cabo el análisis y la solución propuesta, se utilizaron las siguientes fuentes de datos:

data.dictionay.xlsx: Proporciona una descripción detallada de los campos en las bases de datos, facilitando la comprensión de la estructura de datos.

employee_survey_data.csv: Contiene resultados de encuestas de satisfacción laboral de los empleados, esencial para comprender los factores relacionados con los altos retiros.

general_data.csv: Contiene información general de los empleados, como detalles personales, antigüedad y otros datos relevantes para el análisis de retiros.

in_time.csv y **out_time.csv**: Registran la hora de ingreso y salida de los empleados, datos cruciales para evaluar asistencia y puntualidad, relacionados con la satisfacción y retención.

manager_survey_data.csv: Contiene resultados de encuestas de desempeño de empleados realizadas por supervisores, proporcionando percepciones sobre el desempeño de los empleados.

retirement_info.csv: Detalla la información sobre empleados que dejaron la empresa, crucial para comprender patrones y razones detrás de los retiros y su impacto económico.

Los datos se extrajeron con cortes específicos:

La mayoría de los datos se extrajeron hasta el 31 de diciembre de 2015, excepto la información de retiro (retirement_info.csv).

La información de retiro se registró con un corte adicional al 31 de diciembre de 2016 para conocer retiros específicos durante ese año.

Es importante considerar estas fechas al realizar análisis y evaluaciones, ya que las tendencias pueden variar con el tiempo, influyendo en las estrategias de retención propuestas.

Análisis exploratorio

Descripción de los Datos

Para el proceso de limpieza y transformación de los datos se cargaron las bases de datos suministradas por la empresa, haciendo los siguientes procesos en cada una de las bases suministradas:

Información de las columnas: Se analizó todas las columnas y los respectivos datos únicos de cada dataframe, con el fin de analizar cuales datos conformaban cada columna y si tenían datos mal redactados o nulos.
A partir de esta información se encontró que hay variables las cuales no aportaban información relevante, ya que o solo tenían un valor o la información ofrecida no aportaba a la solución del problema. Del mismo modo, se eliminaron variables que pertenecen únicamente a los

empleados despedidos o que han renunciado, debido a que estas no permiten ser evaluadas en el modelo de predicción.

Imputación de datos nulos: Se realizó un análisis de los datos nulos, debido a que los datos faltantes son propios de cada empleado y la mayoría de estos son categorías o está relacionado con el número de compañías donde ha trabajado y el total de años trabajados, se procedió a imputarlos con la moda, con el fin de que no se vean afectados por la variabilidad de estos.

Análisis y corrección del tipo de dato: se procedió a analizar el tipo de datos de cada dataframe y a corregir su naturaleza con ayuda de la base datos "data.dictionay.xlsx". En el caso de las variables categóricas, que están expresadas con números enteros del 1-4 o 1-5, se asignaron como enteras -dato tipo 'int'- con el fin de no perder interpretabilidad; en el caso de las categóricas escritas, se asignaron como objects.

Igualmente, se realizó un tratamiento de datos diferente a las bases de datos "in_time.csv" y "out_time.csv" ya que esta base de datos se conforma de datos de tiempo y, como se mencionó anteriormente, estas tienen el registro de la hora de entrada y salida de cada empleado respectivamente. El tratamiento que se le hizo a estas bases de datos es el siguiente:

Análisis y corrección del tipo de dato: A la hora de analizar los datos de las fechas se encontraba que estaban siendo interpretados como object, por lo que se procedió a darles un formato fecha. Además, los datos nulos no se trataron ya que habían muchos de días festivos o de días los cuales faltaban los empleados, por lo que tratar estos datos era complicado por su naturaleza y también porque esto causa sobreajustes en el modelo predictivo.

Operación entre dataframes: Una vez corregido el tipo de dato se procedió a restar los datos de salida del empleado (out_time) con los datos de entrada (in_time), con el fin de poder obtener el número de horas que había trabajado cada empleado en este año 2015, luego se realizó el promedio de las horas trabajadas de cada empleado, sacando así el promedio de horas que cada trabajador hizo este año -para el promedio de tiempo, se tuvo en cuenta que los datos nulos estaban fuera de los cálculos-.

Exploración de Datos

Durante el proceso de exploración de datos, se hizo un análisis de outliers con la finalidad de conocer qué tantos datos de las variables estudiadas resultan ser datos atípicos, debido a que estos valores extremos pueden afectar la variable objetivo, que en este caso es "Attrition. En dicha exploración, se observó que las variables "MonthlyIncome", "TotalWorkingYears", "YearsAtCompany" y "YearsSinceLastPromotion" presentan un gran número de valores atípicos, pero por la naturaleza de dichos datos no se hace imputación de ninguno de estos valores extremos.

De igual manera, se hicieron boxplots comparativos con la finalidad de ver cómo se comportaban las variables numéricas con la variable respuesta, encontrando como anteriormente, datos atípicos que no se trataron por la naturaleza de dichas variables.

Finalmente, se hizo un uso de histogramas para ver cómo se comporta la variable respuestas con las variables categóricas, encontrando en todas que hay un desbalance de datos, lo que puede causar problemas como el sobreajuste, dificultad para la generalización, sesgos en la clasificación, recall bajo y una precisión engañosa; pero por la naturaleza de la variable respuesta, es normal que en una empresa haya un mayor número de trabajadores contratados que no contratados.

Estrategias de Reducción de Retiros

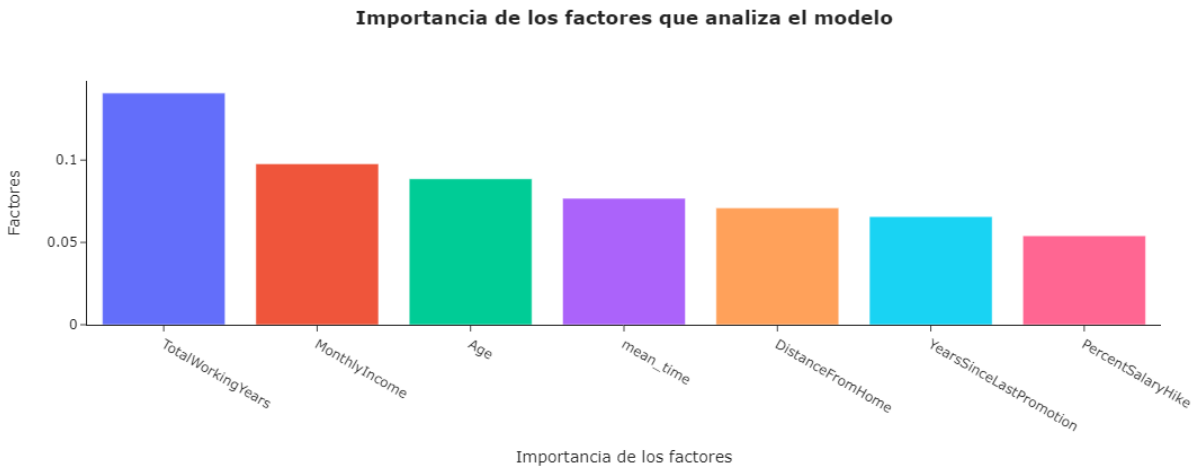
Análisis de Causas

Para el estudio del problema, se definió varios modelos de clasificación, incluyendo la Regresión Logística, Árbol de Decisión, Bosque Aleatorio y Gradient Boosting con la finalidad de identificar las variables más relevantes para el modelo de clasificación. Cada uno de estos modelos, fue evaluado con todas las variables y también con dos métodos de selección de variables los cuales fueron el método Lasso y un método de threshold donde cuya importancia sea mayor que 1.3 veces la media de las importancias de todas las variables.

Para escoger el modelo a trabajar se utilizó la métrica de precisión accuracy, teniendo como resultado que los modelos de mayor predicción son el Decision Tree y el Random Forest con selección de variables por el método treshold. Para efectos prácticos e interpretabilidad del modelo, es escogió el Decision Tree.

Una vez escogido el Decision Tree como modelo predictivo, se realiza una búsqueda de hiperparámetros con la finalidad de probar diferentes combinaciones hiperparámetros para encontrar la configuración que mejor se ajuste al conjunto de datos

Finalmente, se evaluó cuáles eran las variables de mayor influencia en el despido y renuncia de trabajadores, las cuales son las que se presentan en la siguiente imagen.



Evaluación y análisis del modelo

Al momento de hacer la evaluación y análisis del modelo, se encontró que para el conjunto de datos train, este tenía un alto de desempeño debido a que su métrica de desempeño (f1-score)

era de 0.98. Este valor tan alto, posiblemente sea explicado por el hecho de que no se eliminaron datos atípicos y por también por el desbalance en los datos.

Despliegue del modelo

1. **Reentrenamiento semestral del modelo:** Se actualizó el modelo cada seis meses con los datos más recientes de los empleados, ajustando sus hiperparámetros para que refleje las condiciones cambiantes en la empresa.
2. **Indicador trimestral de retiros:** Se establece un KPI trimestral para medir la tasa de retiros, permitiendo una evaluación constante del impacto de las estrategias de mejora.
3. **Meta de reducción anual:** Se fija una meta anual de reducción de retiros de al menos 2%, proporcionando una dirección clara hacia la mejora continua.
4. **Colaboración con empleados:** Los empleados participan activamente en la identificación de estrategias de mejora a través de encuestas y retroalimentación, asegurando que sus perspectivas sean consideradas.
5. **Evaluación de estrategias:** Se analiza el impacto de las estrategias propuestas por el equipo de recursos humanos y los empleados, utilizando métricas específicas para evaluar su efectividad

Al implementar este enfoque completo y cíclico, la empresa estará en una posición más sólida para retener a sus empleados y alcanzar sus objetivos de retención a largo plazo. Además, la retroalimentación continua de los empleados y la evaluación de estrategias permitirán una mejora constante en las prácticas de retención de la empresa.