ANALÍTICA EN RECURSOS HUMANOS

Análisis y Estrategias para Reducir la Tasa de Retiro en una Empresa de 4000 Empleados

UNIVERSIDAD DE ANTIOQUIA
1803

JAVIER BURGOS
CRISTHIAN ALEJO
SUSANA BARRIENTOS

Contenido

Lista de tablas

Lista de ilustraciones

RESUMEN

Este informe presenta un análisis de la deserción de empleados en una empresa específica, utilizando modelos de aprendizaje automático (Machine Learning) para identificar las variables que influyen en la decisión de un empleado de abandonar la empresa. **El objetivo de este trabajo no es identificar a los empleados que están en riesgo de desertar**, ya que esto sería poco ético. En cambio, se busca proporcionar a la empresa información valiosa sobre las variables que están relacionadas con la deserción, para que pueda tomar medidas proactivas para reducirla. ***Ilustración 1***

1. LIMPIEZA Y TRANSFORMACIÓN

Inicialmente se identifica la información brindada en las bases de datos.
Tabla 1

En primer lugar, se lee cada archivo CSV y se filtra el DataFrame de retiros para sólo incluir el año 2016, esto con el sentido de poder predecir el año siguiente (2017), ya que con estos datos del 2016 al ser deserciones efectivas era suficiente data para trabajar los modelos predictivos. Luego, se eliminan filas duplicadas, columnas sin nombre, fechas de encuestas y variables con valores idénticos en todos los DataFrames. Se rellenan los valores faltantes en algunas columnas con el valor 0. Luego, se combinan los 4 DataFrames en uno solo utilizando la columna 'EmployeeID' como referencia, se rellenan los valores faltantes en las columnas 'retirementDate' y 'resignationReason', se convierte la variable 'Attrition' a binaria y se cambia el tipo de la columna 'EmployeeID' a cadena de texto.

El resultado final es un DataFrame limpio y preprocesado listo para ser utilizado en el análisis de la rotación de personal. Se exporta a un archivo CSV llamado 'df_final.csv'.

2. ANÁLISIS EXPLORATORIO

Se exploran las distribuciones de las variables, buscando valores atípicos y visualizando su distribución. Se crea una gráfica de series de tiempo para los retiros y se analizan las variables categóricas mediante gráficos de conteo. Seguido de esto se analiza la relación entre la variable objetivo y las variables categóricas, comparando su distribución entre diferentes categorías. Se calculan pruebas de chi-cuadrado para encontrar las asociaciones entre variables categóricas y se visualizan los "valores p" en un mapa de calor,así

como también se calculan las correlaciones entre variables numéricas mediante otra matriz de correlación.

Finalmente, se realiza una selección de características. Se crea un nuevo conjunto de datos con variables codificadas para las características categóricas y se utiliza un método de eliminación recursiva de características para seleccionar un conjunto final de variables relevantes para el análisis de la rotación de personal

3. SELECCIÓN DE ALGORITMOS Y TÉCNICAS DE MODELADO

La variable respuesta "Attrition" es de particular importancia en la solución del problema, ya que representa una medida binaria que indica si un empleado tiene la tendencia a renunciar en futuros años.

Para abordar el problema se definen los modelos: **Regresión logística (m_lreg)**, **Árbol de decisión (m_rtree)**, **Bosque aleatorio (m_rf)** y **Gradient Boosting (m_gbt)**, ya que estos modelos tienen un buen desempeño frente a variables dependientes que sean binarias permitiendo capturar relaciones complejas entre las características de los empleados y la probabilidad de renuncia. La elección de múltiples modelos proporciona flexibilidad y permite comparar su rendimiento en términos de precisión y capacidad de generalización.

4. SELECCIÓN DE VARIABLES

En la fase de selección de variables, se aplicó una función específica llamada "funciones.sel_variables" la cual permite seleccionar las características más relevantes para un conjunto de modelos de aprendizaje automático anteriormente nombrados. Esta función recibe como entrada los modelos de clasificación a evaluar, las variables independientes (X), la variable objetivo (y) y un umbral de importancia (threshold=2.5*mean).

La función itera sobre cada modelo y realiza un proceso de selección de variables. Primero, ajusta el modelo con las características X y la variable objetivo y. Luego, utiliza la clase SelectFromModel de scikit-learn para seleccionar las características que son más importantes para el modelo, utilizando el umbral threshold como criterio de selección. Finalmente, la función extrae los nombres de las variables seleccionadas y los agrega a un conjunto.

Al final, la función devuelve un conjunto único con los nombres de las variables que fueron seleccionadas por al menos uno de los modelos. Estas variables pueden ser puntos clave para desarrollar estrategias de retención y bienestar laboral**.** ***Tabla 2***

5. COMPARACIÓN Y SELECCIÓN DE TÉCNICAS

El análisis de los resultados obtenidos mediante el modelo K-fold cross validation, enfocado en la evaluación del parámetro F1, revela pocas diferencias en el rendimiento de los modelos al utilizar todas las variables disponibles versus un conjunto reducido de variables seleccionadas. Se utilizó la métrica F1 para evaluar la precisión y la sensibilidad de los modelos en la predicción de la variable objetivo "Attrition". ***Tabla 3***

Los resultados muestran que el desempeño de los modelos varía considerablemente entre ambos escenarios. En el caso de la Regresión Logística, se observa una notable disminución en la puntuación F1 al utilizar las variables seleccionadas, lo que sugiere que estas pueden no ser adecuadas para este modelo en particular. Por otro lado, los modelos de Árbol de Decisión y Bosque Aleatorio muestran un rendimiento excepcionalmente alto en ambas configuraciones, lo que indica una mayor robustez frente a la selección de variables. Sin embargo, el modelo Gradient Boosting también experimenta una disminución en el rendimiento al utilizar variables seleccionadas, aunque no tan pronunciada como en el caso de la Regresión Logística.

La evaluación de la precisión (accuracy) de los modelos proporciona una perspectiva fundamental sobre su desempeño en la predicción de la variable "Attrition". ***Tabla 4***

Los resultados muestran que los modelos de Árbol de Decisión y Bosque Aleatorio logran una precisión excepcionalmente alta en ambas configuraciones. Esto sugiere una capacidad sobresaliente para predecir la rotación de empleados utilizando tanto todas las variables como un subconjunto seleccionado.

6. AFINAMIENTO DE HIPERPARÁMETROS

Se realiza tanto como para el modelo 'Arboles de decisión' como para Random Forest en esta aplicación, se observó que la configuración con el puntaje promedio de prueba más alto fue de (0.961678) coincidiendo conjuntamente con el mejor conjunto de hiperparámetros.

7. EVALUACIÓN Y ANÁLISIS DEL MODELO

Para el modelo Random Forest, las puntuaciones de precisión en el conjunto de entrenamiento son consistentemente altas, alcanzando valores cercanos a 1.0 en todos los pliegues. Esto sugiere que el modelo RF es capaz de ajustarse bien a los datos de entrenamiento. En cuanto al conjunto de prueba, las puntuaciones de precisión también son altas, con una media de aproximadamente 0.961. Esto indica que el modelo RF generaliza bien a datos no vistos y es capaz de mantener un alto rendimiento en diferentes conjuntos de datos.

Para el modelo Árbol de Decisión, las puntuaciones de precisión en el conjunto de entrenamiento son ligeramente más bajas en comparación con el modelo RF, pero aún así son bastante altas. Sin embargo, en el conjunto de prueba, las puntuaciones de precisión son algo más bajas que las del modelo RF, con una media de aproximadamente 0.925. Esto sugiere que el modelo DTree puede estar ligeramente sobreajustado.

8. DESPLIEGUE DEL MODELO

El proceso comienza con la carga del conjunto de datos final (df_final.csv) y la preparación de las características para la predicción mediante la función preparar_datos. Luego, se carga el modelo entrenado de RandomForest(rf_final.pkl) ya que fue este el modelo que tuvo mejor desempeño y se utiliza para realizar predicciones de la variable objetivo (Atrittion) para las nuevas observaciones. Estas predicciones se integran con la información original de los empleados en un nuevo DataFrame (perf_pred).

Para evaluar el rendimiento del modelo, se guarda el DataFrame de predicciones (perf_pred) en un archivo de Excel (prediccion.xlsx). Además, se calcula la importancia de cada característica en la predicción de la variable objetivo y se guarda en un archivo de Excel (importances.xlsx).

Finalmente, se muestran las 10 predicciones de menor probabilidad de deserción (Atrition), ordenadas por la variable objetivo. Esto permite identificar a los empleados con menor riesgo de abandonar la empresa.

Este despliegue facilitara la ejecución del modelo frente a nuevos datos, lo cual permitirá un control del desempeño y de los cambios hechos a partir de el modelo RandomForest inicial, lo cual permitirá una mejora continua en la empresa a la hora de respaldarse en esta herramienta de machine learning

9. CONCLUSIONES:

Los resultados de este análisis proporcionan a la empresa información valiosa sobre las **variables que están relacionadas con la deserción de empleados**. La empresa puede utilizar esta información para tomar medidas proactivas para reducir la deserción, como mejorar la satisfacción laboral, ofrecer oportunidades de desarrollo profesional, promover un mejor equilibrio entre la vida laboral y personal y ofrecer una remuneración competitiva.

Una sugerencia para mejorar la reducción de la deserción en la empresa es fortalecer las iniciativas de bienestar y compensación existentes. Esto puede lograrse mediante la asignación de recursos adicionales, tanto financieros como humanos, para ampliar y mejorar los programas y políticas de bienestar en la organización.

ANEXOS

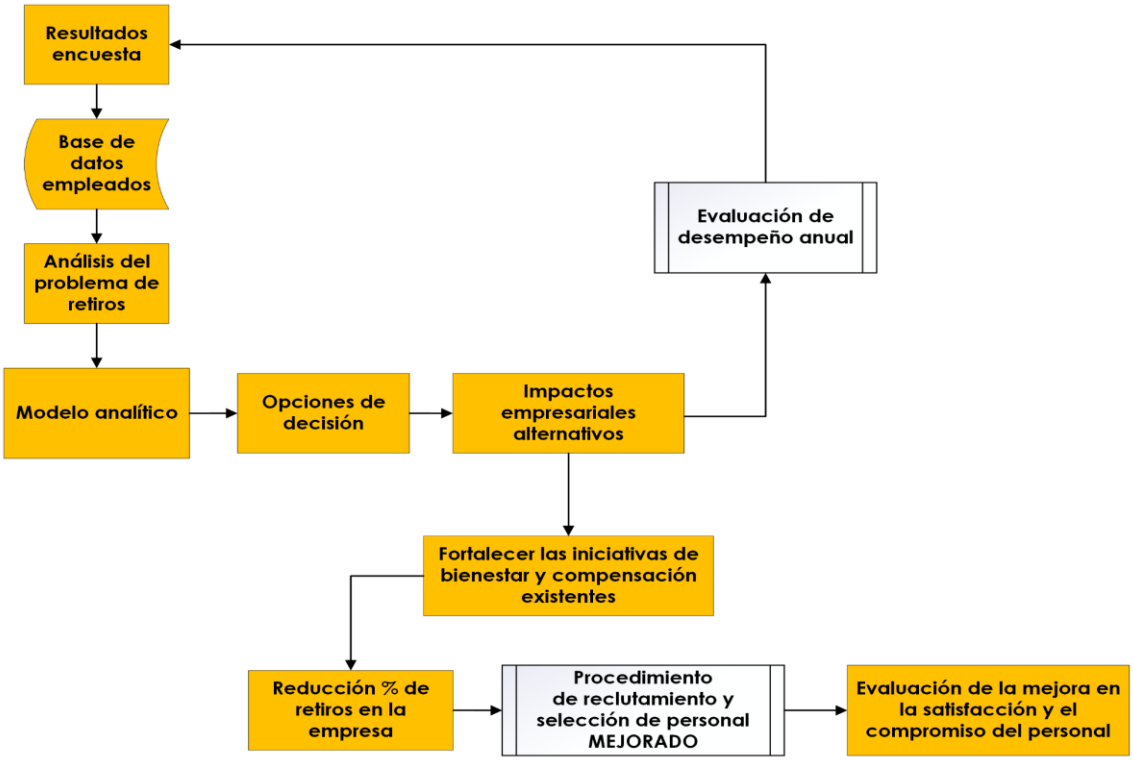

Ilustración 1 Diseño solución propuesta

general_data	Datos generales del empleado: ID, edad, área de trabajo, nivel de educación, género, horas trabajadas, etc.
employee_survey_data	Encuesta realizada a los empleados con variables como: ID, nivel de satisfacción con el entorno de trabajo y con su rol desempeñado, equilibrio entre la vida personal y laboral y fecha en la que se realizó
manager_survey	Encuesta realizada a los gerentes o supervisores con variables como: ID, nivel de participación del gerente, calificación del desempeño del equipo y fecha en la que se realizó.
retirement_info	Información sobre el retiro del empleado con variables como: ID, indicador de retiro de la empresa, calificación de desempeño, día del retiro, tipo y motivo.

Tabla 1 Información Base de datos

Variable	*Peso*
Ingreso Mensual	0,203161
Edad	0,160404
Distancia Desde Casa	0,123295
Años Totales de Trabajo	0,115416
Número de Compañías Trabajadas	0,07925
Años en la Compañía	0,07577
Satisfacción con el Ambiente de Trabajo	0,069356
Satisfacción Laboral	0,058633
Años con el Actual Gerente	0,058377
Años Desde la Última Promoción	0,056338

Tabla 2 Porcentaje de variables

Modelo	F1 (Sin selección)	F1 (con selección)
Regresión Logística	0.227979	0.035714
Árbol de Decisión	0.996923	0.975610
Bosque Aleatorio	1.000.000	0.993865
Gradient Boosting	0.605809	0.446512

Tabla 3 Métricas de desempeño F1 para modelos

Modelo	Precisión (Sin selección)	Precisión (Con selección)
Regresión Logística	0.864914	0.853128
Árbol de Decisión	0.999093	0.990934
Bosque Aleatorio	1.000000	0.999093
Gradient Boosting	0.913871	0.892112

Tabla 4 Métricas de desempeño Accuracy para modelos