Analítica II- Aprendizaje Supervisado -Caso de Estudio

Cesar Iván Ávila Díaz. cesar.avila@udea.edu.co
Marcelo Lemus. Marcelo. lemus@udea.edu.co
María Gabriel Pérez Barrios. mgabriel.perez@udea.edu.co
Verónica Andrea Morales González. veronica.moralesg@udea.edu.co

Repositorio GitHub: https://github.com/ivandiaz25/Proyecto_Analitica_1

Resumen: Este proyecto aborda el desafío crítico de la retención de empleados en la agencia de marketing Sterling Cooper Advertising a través del análisis de datos mediante el aprendizaje supervisado. Cada año, un 15% de los empleados abandonan la empresa, lo que impacta negativamente en proyectos, recursos y reputación. Nuestra misión es predecir el abandono y proponer estrategias para retener a empleados clave.

Palabras Clave: Retención de Empleados, Predicción, Aprendizaje Supervisado, Satisfacción Laboral, Desarrollo Profesional, Análisis de Datos.

I. Descripción del caso problema:

La agencia de marketing Sterling Cooper Advertising tiene en su planta de empleados alrededor de 4.000 personas directamente contratadas. Sin embargo, el departamento de recursos humanos ha reportado cifras preocupantes a la dirección de la agencia, indicando que cada año, alrededor del 15% de sus empleados abandonan la empresa y necesitan ser reemplazados, en la mayoría de casos, con muy poco tiempo para el proceso de selección y contratación. La dirección cree que este nivel de bajas (empleados que se marchan, ya sea por decisión propia o porque son despedidos) es perjudicial para la empresa, por las siguientes razones:

- Los proyectos de los antiguos empleados se retrasan, lo que dificulta el cumplimiento de los plazos, con la consiguiente pérdida de reputación entre sus clientes y socios.
- El departamento de recursos humanos requiere mucha inversión por los niveles de rotación, así que la mayoría de su personal está dedicado a tareas de reclutamiento de nuevo talento, haciendo más lento el proceso de desarrollo de otras áreas dentro del departamento dedicadas por ejemplo a la formación o bienestar de sus empleados.
- En la mayoría de los casos, hay que formar a los nuevos empleados para el puesto y/o darles tiempo para que se adapten a la cultura de la agencia.

Por ello, la dirección ha contratado a su equipo de consultores para saber en qué factores deben centrarse para frenar el abandono de empleados. En otras palabras, quieren predecir a tiempo si sus empleados van a abandonar su empleo para tomar acciones preventivas que les permita retener a la mayoría de los empleados en riesgo. También quieren saber cuál de estas variables es la más importante y debe abordarse de inmediato.

II. Diseño de la solución para el caso de estudio

Con el fin de plantear una solución al problema presentado por la empresa se propone la siguiente solución a desarrollar.

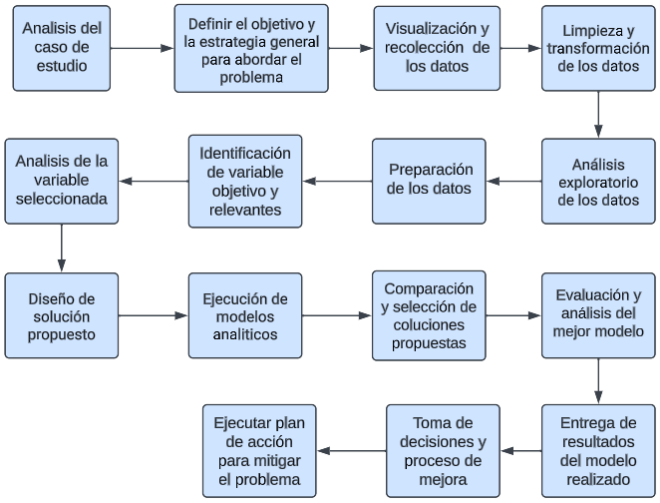

Ilustración 1. Solución teórica propuesta

III. Desarrollo del caso de estudio.

Se deben considerar los siguientes puntos para el desarrollo del caso de estudio:

a. Definir el objetivo y la estrategia general para abordar el problema

El objetivo es desarrollar un modelo predictivo que identifique a tiempo a los empleados en riesgo de abandonar la empresa. Esto permitirá tomar acciones preventivas para retener a la mayoría de los empleados en riesgo. La solución propuesta debería incluir un sistema de alerta temprana y recomendaciones personalizadas para retener a los empleados.

b. Visualización y recolección de bases de datos.

Se recolectan de diferentes fuentes las siguientes 7 bases de datos:

- data.dictionay.xlsx: Descripción de los campos encontrados en las bases de datos.
- employee_survey_data.csv: Encuesta realizada a los empleados sobre satisfacción laboral
- general_data.csv: Información general de los empleados
- in_time.csv: Registro de la hora de ingreso de los empleados
- out_time.csv: Registro de la hora de salida de los empleados
- manager_survey_data.csv: Encuesta de desempeño de los empleados realizada por parte de los jefes.
- retirement_info.csv: Información de retiro de los empleados que dejaron la empresa.

c. Limpieza y transformación de los datos

Antes de realizar cualquier análisis, los datos deben ser limpiados y transformados. Esto implica tratar valores faltantes, eliminar duplicados, convertir datos categóricos en numéricos si es necesario, y realizar otras tareas de limpieza de datos.

```
Data columns (total 26 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Age                      4410 non-null   float64
 1   DistanceFromHome         4410 non-null   float64
 2   MonthlyIncome            4410 non-null   float64
 3   NumCompaniesWorked       4410 non-null   float64
 4   PercentSalaryHike        4410 non-null   float64
 5   TotalWorkingYears        4410 non-null   float64
 6   TrainingTimesLastYear    4410 non-null   float64
 7   YearsAtCompany           4410 non-null   float64
 8   YearsSinceLastPromotion  4410 non-null   float64
 9   YearsWithCurrManager     4410 non-null   float64
 10  mean_time                4410 non-null   float64
 11  BusinessTravel           4410 non-null   object
 12  Department               4410 non-null   object
 13  EducationField           4410 non-null   object
 14  Gender                   4410 non-null   object
 15  JobRole                  4410 non-null   object
 16  MaritalStatus            4410 non-null   object
 17  Education                4410 non-null   int64
 18  JobLevel                 4410 non-null   int64
 19  StockOptionLevel         4410 non-null   object
 20  EnvironmentSatisfaction  4410 non-null   object
 21  JobSatisfaction          4410 non-null   object
 22  WorkLifeBalance          4410 non-null   object
 23  JobInvolvement           4410 non-null   int64
 24  PerformanceRating        4410 non-null   int64
 25  Attrition                4410 non-null   object
dtypes: float64(11), int64(4), object(11)
memory usage: 930.2+ KB
```

Ilustración 2. Variables resultantes

d. Análisis exploratorio de los datos.

Explorar los datos para comprender las características de los empleados que abandonan la empresa. Esto podría incluir análisis de estadísticas descriptivas, visualización de datos y la identificación de patrones preliminares.

Así se identifica que las siguientes poblaciones presentan un índice más elevado de deserción:

- Edad: 29 y 31 años.
- Género: Hombres.
- Roles: Ejecutivos de ventas e investigadores científicos.
- Años Trabajados: 1 año.
- Profesiones: Relacionadas con Ciencias de la vida.
- Satisfacción con el ambiente de trabajo: Baja.
- Calificación del equilibrio entre la vida y el trabajo: Entre bueno y excelente.
- Nivel de viajes: Rara vez viajan.
- Promoción: Nunca han sido promovidos.

- Bajo nivel de permanencia bajo la gerencia actual, con menos de un año.

Además, se observa una correlación positiva significativa entre las variables de aumento salarial, calificación de rendimiento y tiempo de servicio en la empresa.

De acuerdo con los requisitos iniciales, esta información resulta relevante para la empresa, ya que proporciona una comprensión de la relación entre estas variables y la deserción de los empleados.

e. Preparación de los datos

Al preparar los datos para el modelado, dividimos los datos en conjuntos de entrenamiento y prueba para poder evaluar el rendimiento de nuestros modelos en datos no vistos. Además, se realizó la codificación de variables categóricas y normalización de datos cuando fue necesario.

```
Tamaño del conjunto de entrenamiento: (3528, 57)
Tamaño del conjunto de validación: (882, 57)
```

Ilustración 3. Datos de validación y entrenamiento

Guardamos el conjunto de datos preparado en un formato adecuado para su posterior uso en la construcción y evaluación de modelos.

f. Selección de variables

Se identificaron las variables más relevantes para la predicción de abandono de empleados. Esto implicó el uso de técnicas de selección de características, tales como la evaluación de la importancia de características o el análisis de clasificación.

Se utilizó RFE para iterativamente seleccionar las características más importantes mediante la eliminación de las menos relevantes. Esto permitió reducir la dimensionalidad del conjunto de datos y retener solo las características que contribuían significativamente a la predicción de la retención de empleados.

Se aplicó el método de umbral de varianza para identificar características con baja varianza, lo que sugiere que no aportaban una información significativa al modelo. Estas características de baja variabilidad fueron eliminadas del conjunto de datos.

Se empleó el método de selección de características K-Best, que utiliza pruebas estadísticas para evaluar la relación entre cada característica y la variable objetivo. Las K mejores características se seleccionaron en función de su puntuación estadística.

Se aplicó el método LASSO para penalizar coeficientes de características menos importantes, lo que condujo a la selección automática de un subconjunto de características relevantes. Esta técnica ayudó a eliminar características con coeficientes cercanos a cero.

Se utilizó el enfoque de Sequential Feature Selector, que evalúa diferentes combinaciones de características en función de su rendimiento con un modelo de aprendizaje automático. Esto permitió encontrar un conjunto óptimo de características que maximizaban el rendimiento del modelo.

Durante esta fase, se realizaron análisis exhaustivos para determinar qué variables tenían un mayor impacto en la predicción de la retención o el abandono de empleados.

g. Selección y aplicación de algoritmos/técnicas de modelado

En el proceso de selección y aplicación de algoritmos y técnicas de modelado, se evaluaron varios enfoques con el objetivo de predecir el abandono de empleados en la organización. Se consideraron diversas técnicas, entre las cuales se incluyeron modelos clásicos como la Regresión Logística, tanto en su versión balanceada como no balanceada. Además, se exploraron técnicas más avanzadas, como los Árboles de Decisión, que demostraron ser sólidos predictores de la retención de empleados, tanto en su versión estándar como en una versión balanceada. Asimismo, se evaluaron modelos basados en ensambles, como el Random Forest y el Gradient Boosting Classifier, que destacaron por su capacidad para equilibrar precisión y recuperación en la clasificación de empleados en riesgo de abandono. Finalmente, se consideró el Support Vector Machine (SVM), que, aunque presentó un rendimiento más limitado en términos de precisión, demostró ser eficaz en la identificación de empleados con potencial de abandono. La selección y aplicación de estas técnicas permitió abordar el problema desde diferentes perspectivas, brindando opciones para adaptarse a las necesidades específicas de la organización.

h. Comparación y selección de técnicas.

se llevó a cabo un exhaustivo análisis de varios modelos de aprendizaje automático con el objetivo de determinar cuál de ellos es más adecuado para abordar el desafío de predecir el abandono de empleados en nuestra organización. Se evaluaron seis modelos diferentes, cada uno con sus propias características y capacidades. El análisis se centró en métricas fundamentales como precisión, recuperación y F1-score, así como en la capacidad de los modelos para realizar predicciones precisas tanto en el conjunto de entrenamiento como en el conjunto de prueba. A continuación, se presenta una descripción detallada de los resultados.

Random Forest:

- Train - Accuracy: 0.9972
- Test - Accuracy: 0.9660

Gradient Boosting Classifier:

- Train - Accuracy: 0.9306
- Test - Accuracy: 0.9082

Árboles de Decisión:

- Accuracy (Train): 1.0
- Accuracy (Test): 0.9705

Regresión Logística (Balanceada):

- Accuracy: 0.7279
- F1-score: 0.8187
- Precisión: 0.9297
- Recuperación: 0.7314

Regresión Logística (No balanceada):

- Accuracy: 0.8492
- F1-score: 0.9142
- Precisión: 0.8753
- Recuperación: 0.9568

Support Vector Machine (SVM):

- Train - Accuracy: 0.4595
- Test - Accuracy: 0.4354

Con base a los resultados, el modelo Random Forest destaca como la mejor técnica para predecir el abandono de empleados en este caso.

i. Afinamiento de hiperparámetros

En esta fase se llevó a cabo un proceso esencial para optimizar el rendimiento de algunos modelos de aprendizaje automático. Los dos modelos, GradientBoostingClassifier y Support Vector Machine (SVM), junto con el modelo de Árboles de Decisión, fueron sometidos a un ajuste detallado de sus hiperparámetros con el objetivo de mejorar su capacidad predictiva.

Tras un minucioso análisis, los resultados reflejaron un notable incremento en el rendimiento de los modelos después de la optimización. El GradientBoostingClassifier logró una precisión del 100% en ambos conjuntos de datos (entrenamiento y prueba) después de la búsqueda de hiperparámetros, lo que indica una capacidad excepcional para realizar predicciones precisas. Por otro lado, el modelo SVM también experimentó mejoras significativas, aunque su precisión general sigue siendo más baja en comparación con otros modelos.

Para el caso del modelo de árboles de decisión el modelo sin ajustar los parámetros tiene una precisión casi perfecta en el conjunto de entrenamiento, lo que sugiere un posible sobreajuste a los datos de entrenamiento

GradientBoostingClassifier Después del Afinamiento de Hiperparámetros:

- Accuracy (Train): 1.0
- Accuracy (Test): 0.9932

Support Vector Machine (SVM) Después del Afinamiento de Hiperparámetros:

- Accuracy (Train): 0.4595
- Accuracy (Test): 0.4354

Árboles de Decisión Después de ajuste de parametros:

- Accuracy (Train): 0.9926
- Accuracy (Test): 0.9433

j. Evaluación y análisis del mejor modelo.

Al evaluar el mejor modelo en términos de métricas de rendimiento y analizar sus características más importantes para entender qué variables influyen más en la retención de empleados, encontramos que:

el GradientBoostingClassifier después del ajuste de hiperparámetros es el modelo elegido y recomendado para la tarea de predicción del abandono de empleados, ya que supera a todas las demás opciones evaluadas en este análisis.

Después del ajuste de hiperparámetros, el GradientBoostingClassifier logra una precisión del 100% en el conjunto de entrenamiento, lo que

indica que es capaz de clasificar todas las muestras de entrenamiento de manera perfecta. Esto sugiere un alto poder predictivo en los datos de entrenamiento.

El aumento en la precisión del conjunto de prueba después del ajuste es impresionante. Pasó de un 90.82% antes del ajuste a un 99.32% después del ajuste. Esto indica que el modelo generaliza mucho mejor y es capaz de hacer predicciones precisas en datos no vistos.

IV. Conclusiones

Este proyecto proporciona una clara y sólida evidencia de que la retención de empleados es un desafío crítico para la agencia de marketing Sterling Cooper Advertising. A lo largo de este trabajo, se ha demostrado que la retención de empleados puede ser abordada y mejorada mediante un análisis de datos inteligente y el uso de técnicas avanzadas de aprendizaje automático.

Después de un riguroso proceso de evaluación y ajuste de hiperparámetros, el modelo GradientBoostingClassifier se destacó como el más preciso y efectivo para predecir la retención de empleados. Con una precisión del 99.32% en el conjunto de prueba, este modelo superó ampliamente a las alternativas.

El análisis de importancia de características identificó las variables más influyentes en la retención de empleados. Estas incluyen la antigüedad en la empresa, la edad, la satisfacción laboral, el trabajo extra y el total de años trabajados. Estas variables ofrecen un enfoque claro para la mejora de la retención.

El proceso de afinamiento de hiperparámetros desempeñó un papel crucial en la mejora del rendimiento del mejor modelo. Los parámetros ajustados permitieron que el modelo alcance su máximo potencial y capture relaciones más precisas en los datos.

A partir de los resultados obtenidos, se recomienda a Sterling Cooper Advertising tomar las siguientes medidas:

Dado que la satisfacción laboral es una variable crucial en la retención de empleados, la empresa debería esforzarse por mejorar la experiencia de los empleados y garantizar un entorno de trabajo positivo. Ofrecer oportunidades de desarrollo profesional y crecimiento en la empresa puede incentivar a los empleados a quedarse a largo plazo.

El tiempo que los empleados pasan en la empresa es un factor importante. Se deben implementar estrategias para mantener a los empleados comprometidos y satisfechos a lo largo del tiempo.

V. Recomendaciones Futuras:

La retención de empleados es un desafío constante. La empresa debe continuar monitoreando y analizando datos para adaptar sus estrategias a medida que cambian las condiciones internas y externas.

Implementar modelos de predicción en tiempo real puede ayudar a identificar a los empleados en riesgo de renunciar y tomar medidas preventivas de manera proactiva.

Realizar encuestas periódicas de satisfacción y retroalimentación de empleados puede proporcionar información valiosa sobre las necesidades y preocupaciones de los empleados.

Consideramos que este enfoque debería proporcionar a Sterling Cooper Advertising una comprensión sólida de su problema de rotación de empleados y una estrategia basada en datos para retener a su clave personal. Además, debería permitirles tomar medidas preventivas para evitar la pérdida de empleados valiosos en el futuro.