UNIVERSIDAD DE ANTIOQUIA

Aplicaciones de la Analítica - Analítica en Recursos Humanos

Alanis Álvarez, Juan E. Cardona, Juan E. Soto, Santiago Restrepo
Estudiantes de Ingeniería Industrial
Universidad de Antioquia, Colombia

1. Diseño de la solución

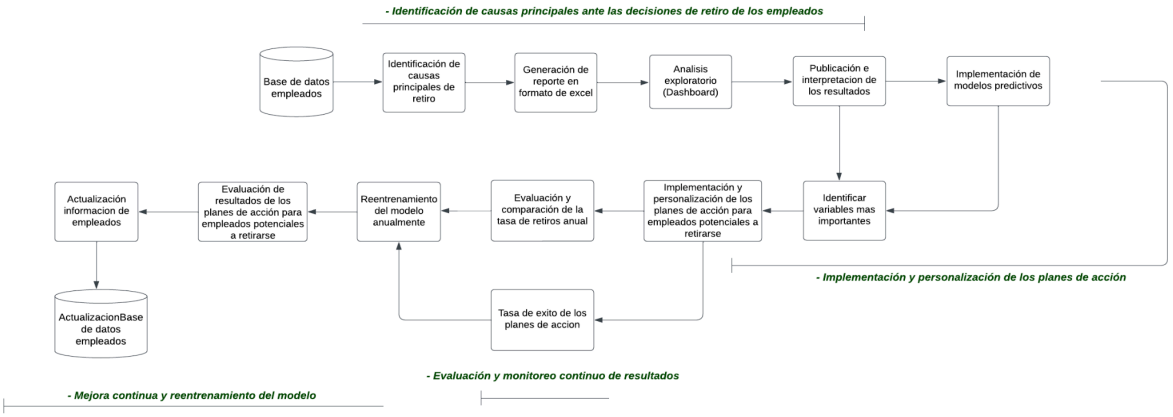

Imagen 1. Diagrama del diseño de la solución caso Recursos Humanos
Fuente. Elaboración propia en Lucidchart

- Identificación de causas principales ante las decisiones de retiro de los empleados

Se generará un reporte en formato de excel que incluirá el ID de los empleados propensos a retirarse y las variables más importantes que influyen en su decisión. Este reporte será complementado con un Dashboard interactivo en Power BI, el cual permitirá a los jefes de área y a los encargados en Recursos Humanos visualizar rápidamente de manera dinámica el comportamiento de las variables clave, visualizando las diferentes causas de retiro y así atacar directamente las causantes de su decisión.

- Implementación y personalización de los planes de acción

Una vez que se han identificados los empleados potenciales a retirarse analizados tanto en el reporte de Excel como en el Dashboard, el área de Recursos Humanos podrá proceder a la implementación de acciones personalizadas con el fin de atacar las causas principales que pueden llevar a un empleado a retirarse. Ahora bien, como todos los planes de acción serán específicos y personalizados para cada empleado, se aseguraría que se aborden las razones subyacentes detrás de la predicción de retiro, con el fin de mantener altas probabilidades de retención de empleados en la empresa.

- Evaluación y monitoreo continuo de resultados

Es fundamental realizar seguimiento constante, para ello, se evaluará la tasa de retiros de empleados cada año. Para ello, se compara la tasa de deserción inicial (15%) con la tasa obtenida en cada evaluación, permitiendo verificar si los planes de acción si están contribuyendo a la reducción de la tasa de retiro, y en caso de que no esté funcionando, se pueda ajustar o rediseñar las estrategias personalizadas.

- Mejora continua y reentrenamiento del modelo

Finalmente, cada año el modelo predictivo será entrenado con los datos actualizados para mejorar su precisión y adaptarse a cambios en los patrones de retiro de los empleados, esto quiere decir que el Dashboard también va a continuar en constante actualización para que se continue dinamizando y facilitando la interpretación, además se va a garantizar que el análisis se mantenga al día y que los planes de acción sigan respondiendo de manera efectiva.

2. Preprocesamiento

Primeramente, después de tener consolidadas las bases de datos, se identifica que las variables *'EmployeeCount', 'Over18'* y *'Standard Hours'* son variables que cuentan con un único valor, por ende se decide eliminarlas, ya que no aportan información relevante. Por otra parte, la variable *'JobRole'* contiene 9 categorías diferentes, las cuales estaban contenidas dentro de la variable *'Department'*, por esta razón, para evitar aumentar la dimensión del dataframe se decide eliminar *'JobRole'*. También se elimina la variable *'Gender'*, porque la proporción de hombres o mujeres es muy similar a la que se retira por lo que no ofrece una explicabilidad significativa y finalmente, la variable *'PerformanceRating'*, solo posee 2 valores, (3 y 4) los cuales son valores muy cercanos por lo que no permite una diferenciación que sea útil.

3. Análisis exploratorio

Se realizó la matriz de correlación, donde se identificó una alta correlación entre las variables '*Age*' y '*TotalWorkingYears*', esto tiene sentido ya que entre más años tenga un empleado, más es el total de años trabajados dentro de la empresa. También se evidenció una alta correlación entre '*YearsAtCompany*' y '*YearsWithCurrManager*', en el momento no es relevante ya que en el análisis se tiene de línea temporal un año, al igual que las variables '*YearsSinceLastPromotion*' y '*YearsAtCompany'*. Entonces se eliminaron las variables de *'TotalWorkingYears', 'YearsAtCompany'* y *'YearsWithCurrManager*'. Para la variable objetivo 'Attrition' es evidente el desbalance entre las clases en donde la cantidad de empleados que permanecen es del 85.3% contrastado con el 14.7% que se retiraron para el 2016. Con respecto a la edad, se evidenció que los empleados que renuncian tienen mayor concentración de los datos entre los 25 y 35 años.

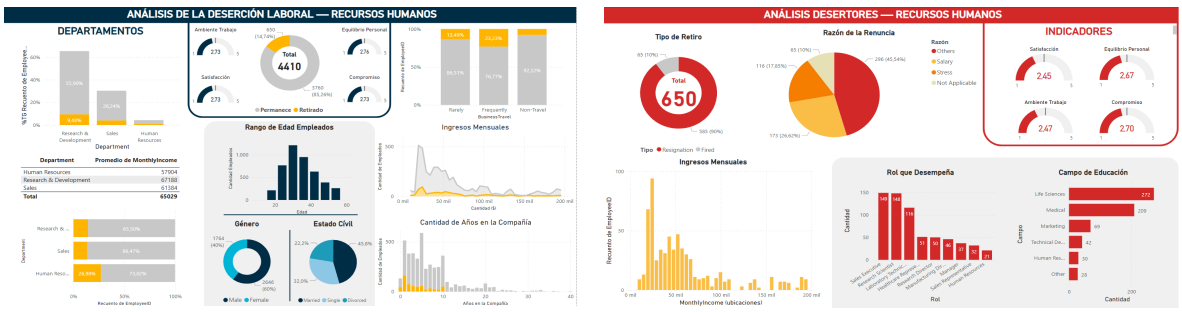

Imagen 2. Dashboards sobre los datos de estudio
Fuente. Elaboración propia en Power BI

Finalmente, se ejecutó un análisis gráfico a través de un tablero de Power BI (ver Imagen 2), y gracias a esto, se ejecutan un par de análisis que nos arrojan como resultado que variables como '*Age'* y '*MaritalStatus*' son algunas de las variables que más influyen sobre la decisión del retiro del empleado. Esto, será comprobado a través de modelos de Shallow Learning que se detallan más adelante.

4. Selección de variables

Adicionalmente, se obtienen los datos resultantes del análisis exploratorio y se evidencia que las variables *'retirementType'* y *'resignationReason'* arrojan la misma información de la variable objetivo, por ende se eliminan del dataframe dado que generaría un sobreajuste en el comportamiento de los modelos. Adicionalmente, estos datos se llevan a formato dummies y se escalan, obteniendo un conjunto de datos con 25 columnas.

Ahora, se realiza la selección de variables obteniendo como resultado un dataframe con 19 columnas. Y para la comprobación de una selección de variables adecuada, se evalúa el conjunto de datos en 4 modelos (se explican más adelante) usando la métrica de precisión.

5. Optimización de hiperparámetros

Al observar el código se evidencia que se ejecuta un proceso de optimización de hiperparámetros a través del método de búsqueda aleatoria para el modelo de Random Forest, el cual logró obtener una mejoría leve en la precisión. Sin embargo, al visualizar la matriz de confusión se evidencia que la tasa de predicción de los verdaderos positivos es muy baja y la predicción de los verdaderos negativos es bastante buena, por lo que el modelo se inclinaba en gran medida por la clase mayoritaria (aún así usando *class_weight = 'balanced'*). Es por eso entonces que se decide continuar haciendo uso del modelo de Random Forest sin optimización de hiperparámetros.

6. Modelos

Para este proyecto se decide optar por el uso de algoritmos (modelos) de Shallow Learning. En este caso, es útil el uso de estos modelos dado que, por la naturaleza de los datos, se cuenta con una variable objetivo en train que puede ser aprovechada para obtener resultados y predicciones más significativas que en el caso de un método de clusterización. Además, no se hace uso de deep learning dado que se considera que los datos no son tan complejos ni extensos para optar por esa opción.

Por otro lado, se seleccionan 4 modelos candidatos para generar las predicciones (logistic regression, decision tree, random forest y gradient boosting machine), de este modo, se cuenta con la versatilidad de los árboles de decisión y la regresión logística acompañada de la robustez del random forest y gradient boosting. Adicionalmente, son modelos que a pesar de ser propensos al sobreajuste, se pueden tratar para evitarlo y obtener una excelente clasificación.

Adicionalmente, se utiliza la **precisión** como métrica de desempeño haciendo uso de los 4 modelos mencionados. Esto, dado que en la problemática actual es crucial evitar la presencia

de falsos positivos, ya que esto incurriría en costos de implementación de planes de acción con empleados que no lo requieren. Al mismo tiempo, se busca maximizar la tasa de verdaderos positivos y esta métrica nos permite dar una muestra del comportamiento de estos. Finalmente, haciendo uso de esta métrica, el modelo que presentó una precisión mayor fue Random Forest, el cual utilizando cross validation arroja una precisión cercana al 96% en entrenamiento y del 93,1% en validación, lo que permite concluir que las predicciones generadas son muy precisas, tomando la decisión definitiva de utilizar el modelo de Random Forest sin optimización de hiperparámetros y utilizando el conjunto de datos con selección de variables para generar las predicciones de 2017. A continuación se visualiza la matriz de confusión asociada a los resultados del modelo (ver Imagen 3). Dentro de la matriz se evidencia que el modelo cuenta con una tasa de verdaderos positivos notable, y que al mismo tiempo, predice bastante bien los verdaderos negativos mientras se reduce la tasa de falsos negativos y falsos positivos.

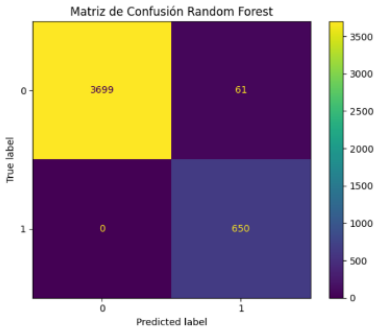

Imagen 3. Matriz de confusión Random Forest
Fuente. Elaboración propia en Python

Finalmente, se imprime el peso de las variables en la predicción del modelo, obteniendo como resultado que las principales variables que influyen en la predicción son '*Age*' y '*MonthlyIncome*'. Así mismo, variables relacionadas con la satisfacción, el campo de educación y el estado civil también influyen sobre la predicción pero en menor medida. Esto permite confirmar los análisis realizados en Power BI con respecto a las variables que consideraba más importantes.

7. Despliegue

Ahora, para generar la predicción de los retiros para el año 2017, se hace uso entonces de la base de datos con los empleados del 2016. Una vez realizada la predicción se obtiene como resultado que el 15.56% de los empleados se retirarán (lo que equivale a 686 empleados).

Adicionalmente, se realiza un análisis de los empleados que el modelo predice como retiros bajo algunas variables importantes como la edad, el estado civil y el sueldo mensual. Gracias a esto, se evidencia que los empleados que se retiran son mayormente solteros y su edad en promedio es de 36 años, lo que permite apreciar (gracias al análisis exploratorio) que los empleados que se retiran suelen ser solteros y un poco más jóvenes que los que se mantienen. Así mismo, el sueldo mensual promedio de estos empleados es de $61.800, que es ligeramente menor al sueldo de los empleados que se mantienen.

8. Recomendaciones y conclusiones

Entonces, una vez identificados los empleados que se retirarían de la empresa y las variables más importantes, se busca reducir la tasa de deserción a través de la puesta en marcha de los planes de acción. Así, insistiendo en que las variables más relevantes son la edad y el ingreso mensual se proponen las siguientes estrategias específicas para atacar dichos retiros y poderlos evitar:

- **Bonos por antigüedad:** El objetivo de estos bonos es que si un empleado cumple cierta cantidad de años de antigüedad en la empresa, este pueda reclamar su bono ya sea en tiempo o monetario, acompañado de un reconocimiento por su labor.
- **Bonos por cumplimiento de metas:** Este bono sirve como incentivo individual y consiste en que el empleado recibirá un bono monetario sujeto al cumplimiento de una meta definida por su superior. Así, el empleado trabajará en pro de aumentar su ingreso mensual y al mismo tiempo aumentará la eficiencia en su cargo.
- **Beneficios y desarrollo profesional:** Dado que las personas están en constante aprendizaje estando en un entorno competitivo sería útil contar con convenios con diferentes plataformas o entes educativos con descuentos que le permitan acceder a diferentes capacitaciones o certificaciones, y de esa forma aumentar su salario. Así mismo, para atacar la edad como motivo de retiro se propone implementar descuentos en actividades que llaman la atención a la población más joven como el gimnasio, créditos de vivienda y hasta la posibilidad de trabajar remotamente.
- **Crecimiento dentro de la empresa:** Desarrollar planes de carrera que le permitan al empleado conocer la manera de aspirar a crecer dentro de la empresa. De esta forma, el empleado será consciente de los requisitos para ascender u ocupar otro cargo, y así, aumentar su valor dentro de la organización y su compromiso con la misma, lo que le permitirá mantenerse por más tiempo en la empresa.

Por otro lado, para evaluar la eficiencia de los planes de acción se define una **tasa de éxito de las intervenciones** realizadas por el área de Recursos Humanos con cada empleado que se retiraría. Esta consiste en evaluar cuántos de los empleados que el modelo predice como retiro cambiaría su decisión después de implementar el plan de acción, y se busca que esta tasa sea lo más cercana a 1. Así mismo, se busca reducir los **costos de reclutamiento y capacitación** de empleados nuevos que reemplazarían a los que se retirarán, y para esto, se podría definir el porcentaje de recursos que la empresa destina a esto, y evaluar si al final del año este valor se redujo respecto al año anterior. De esta forma, se espera que la tasa de deserción sea inferior al 12%.

Finalmente, se cuenta con las siguientes conclusiones:

- La analítica de datos es una herramienta poderosa para realizar un diagnóstico del comportamiento de las empresas. En este caso, el análisis exploratorio y la implementación de modelos nos permitió llegar a definir cuáles son los aspectos que generan la deserción en esta empresa, y de esa forma, realizar recomendaciones orientadas a ello.

- El uso y resultados de los modelos depende notablemente de la estructura y complejidad de los datos. Así mismo, se requiere que los datos contengan las variables adecuadas para su funcionamiento y predicción.

Enlace del repositorio: https://github.com/juanestebansoto/ProyectoRH.git