INFORME
DISEÑO DE LA SOLUCIÓN

Para el diseño de la solución se identificaron los componentes analíticos que permiten predecir las renuncias totales del año siguiente, se realizó una limpieza y transformación de los datos de tal forma que las variables a estudiar fueran representativas para el área de recursos humanos. La métrica analítica empleada fue la F1 Score, una métrica idónea para medir precisión y exhaustividad en los modelos y que para el caso no presenta problemas con las clases desbalanceadas, además, se entrenó con la información del año 2015 para predecir el 2016; luego se utilizaron la información del 2016 para el despliegue y predecir 2017; se estudiarán las renuncias y sus razones de tal forma que permita implementar estrategias de y esta manera reducir la deserción laboral. Por consiguiente se hace el despliegue del modelo en donde se le entregará al área de recursos humanos cuatro tablas, la primera son las variables más influyentes en la deserción laboral, segundo los empleados que renunciaron con sus variables originales, tercero la misma tabla anterior pero con la tabla estandarizada (para poder realizar comparaciones con el umbral de los nodos de las rutas de decisión) y por último la ruta de decisión para cada empleado, todas cuatro en excel y solo las dos primeras en la base de datos DB. Posteriormente realizar una evaluación y análisis de resultados del modelo predictivo de renuncias para el 2017 y que el área logre diseñar los planes para mitigar la problemática.

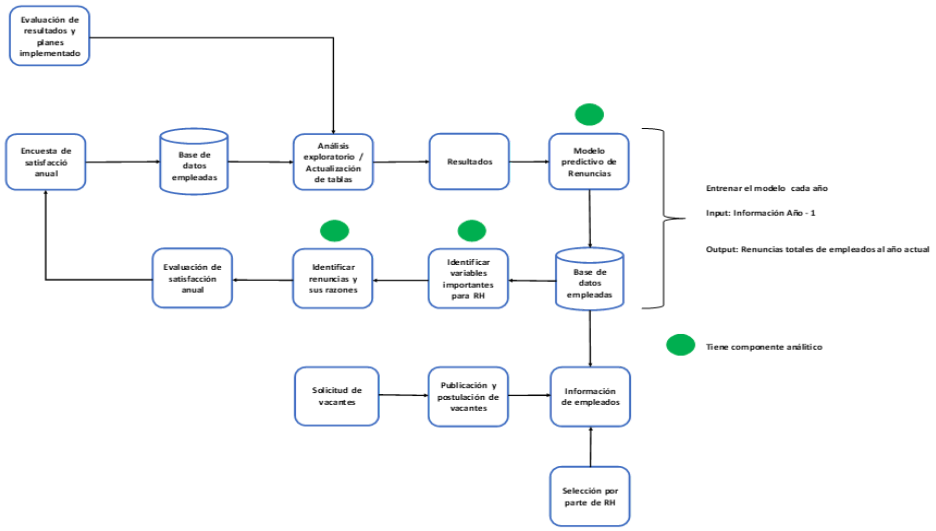


Gráfico 1: Diseño de solución

Limpieza y transformación

Se inició con la lectura y revisión inicial de cuatro conjuntos de datos, identificando variables irrelevantes o constantes para su eliminación dentro de SQL. Se convirtieron las fechas a un formato consistente y se llenaron valores nulos con la mediana en variables clave. A través de un análisis de correlaciones y pruebas chi-cuadrado, se eliminaron variables con correlaciones altas. Finalmente, se prepararon los datos limpios para el análisis exploratorio, exportándolos para su uso en fases posteriores del proyecto.

Análisis exploratorio.

El análisis exploratorio, luego de la limpieza de datos y la transformación adecuada, muestra que no se tienen nulos en la tabla final. Se analizan tanto variables numéricas como categorías, tanto por separado como la relación que tiene cada una con la variable objetivo. Algunos datos relevantes que se pueden extraer de allí bien pueden ser sobre cómo los desertores tenían una peor calificación en las encuestas de satisfacción con respecto a los no desertores. Según el análisis, se logra observar cómo probablemente los desertores tienen en promedio menor edad. De las variables categóricas se tienen observaciones como que las personas que no viajan desertan menos en proporción a las que sí lo hacen; también se tiene que para los médicos y los científicos se puede encontrar mayor proporción de desertores. Y por último, los solteros sí suelen desertar más que los casados y los divorciados según la gráfica.

Selección de variables

En la selección de variables se aplicaron diferentes métodos, el primero de ellos fue el método de filtrado en donde se eliminaron variables como “over18”, “StandardHours”, entre otras, estas eran constantes y no aportan a la explicabilidad de la deserción, además se realizó la prueba chi-cuadrado en donde se eliminó “género” debido a su alta correlación con las demás variables categóricas. También se realizó la matriz de correlación en donde se eliminaron las variables “yearsatcompany”,” totalworkingyears” y ”performancerating”, estas presentaban una alta correlación con otras variables. A partir de los estimadores logistic regression, decision tree classifier, random forest classifier y gradient boosting classifier y con dos threshold de 2.2*mean y 2.5*mean se seleccionaron 8 variables y 5 variables respectivamente, además se realizó un gráfico que itera sobre una lista de thresholds y busca el mejor desempeño en cada modelo, los resultados fueron los siguientes: reg logistic threshold 0.5, decision_tree threshold 4.4, random_forest threshold 3.5 gradient boosting threshold 1.25. La gráfica 2 muestra los resultados obtenidos, cabe resaltar que los threshold para los modelos que obtuvieron inicialmente mejor desempeño (con threshold 2.2*mean) los cuales fueron RFC Y DTC con sus respectivos mejores valores de threshold arrojaban máximo 2 o 3 variables explicativas. Por último como se observa en el gráfico 3 el desempeño con más variables no afectaba estadísticamente los resultados, por lo que se definió un threshold de 2.5 que seleccionara 5 variables explicativas.

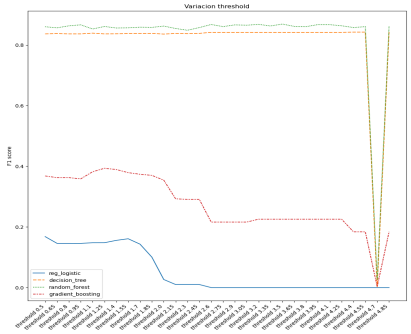


Gráfico 2: Resultados de threshold

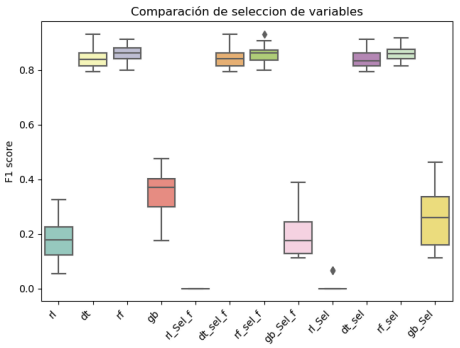


Gráfico 3: Desempeño de las variables

Comparación y selección de técnicas

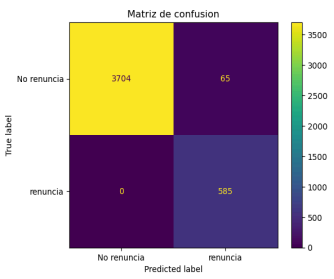
Luego de probar los diferentes grupos de variables: todas las variables, con cinco variables y por último con 8 variables, respectivamente como se puede observar en el gráfico 4 se concluye que el desempeño al eliminar variables en los algoritmos de DTC Y RFC no se ve estadísticamente afectado, por esta razón y en la búsqueda de un modelo con mayor explicabilidad con un menor número de variables se decide continuar con los modelos DTC y RFC. la regresión logística y el gradient boosting classifier si alteran considerablemente su desempeño.

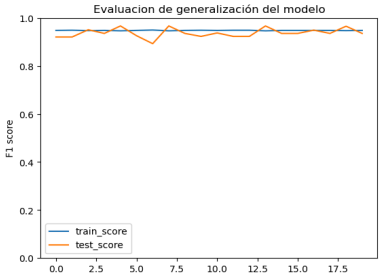
Afinamiento de hiperparámetros

Después del análisis anterior se realiza el tunning de los modelos DTC y RFC, en ambos casos se obtiene una mejora de 84,16% a 93,65% en el DTC 85,82% a 93,71% en el RFC, mejorado considerablemente el f1 score en ambos modelos. Finalmente el modelo escogido es el DTC, ya que la métrica de evaluación es bastante similar al RFC, sin embargo el modelo DTC tiene una mejor interpretabilidad para el análisis de las variables que afectan en la renuncia de un empleado.

Evaluación y análisis del modelo

Luego de escoger el modelo DTC, el modelo está generalizando correctamente la información, se puede ver en el gráfico 5. En el análisis del modelo se tiene que el modelo predice correctamente el 100% de las veces cuando el empleado no va a renunciar y un 90% de las veces acierta cuando va a renunciar sobre las predicciones, además es capaz de predecir el 98% de las personas que no van a renunciar y el 100% de las personas que van a renunciar en realidad. En nuestro problema de negocio un falso negativo tendría un mayor impacto en los resultados, debido a que el modelo no sería capaz de predecir cuándo realmente un empleado renuncia y en tal caso no poder abordarlo a través de planes para prevenir esta decisión, en el caso contrario un falso positivo se perdería el plan o acompañamiento a dicho empleado, aun así esto puede afianzar más el empleado a la empresa.






```
Train - Accuracy : 0.9850711988975654
Train - classification report:
              precision    recall  f1-score   support

           0       1.00      0.98      0.99      3769
           1       0.90      1.00      0.95       585

    accuracy                           0.99      4354
   macro avg       0.95      0.99      0.97      4354
weighted avg       0.99      0.99      0.99      4354
```

Gráfico 4: Matriz de confusión *Gráfica 5: Evaluación de generalización* *Tabla 1: Informe de clasificación*

Despliegue del modelo

Predicciones: Las predicciones serán entregadas en db basedatos llamada prediccion_renuncias que incluye los empleados que desertaran con las variables

más importantes que encontró el modelo, se ingresara la base de datos la información de todo un año (2016) y predice la deserción en el año siguiente (2017)

Importancia de las variables: serán entregadas en db basedatos llamada importancia_variables, estas variables tienen una mayor influencia para la clasificación de los empleados

Resultados:

Variables importantes, predicciones y ruta de decisión:

Feature	Importance
monthlyincome	0,419684509
age	0,248131773
yearswithcurrmanager	0,168493668
yearssincelastpromotion	0,09864398
jobsatisfaction	0,06504607

Tabla 2: Variables más influyentes

age	jobsatisfaction	monthlyincome	yearssincelastpromotion	yearswithcurrmanager	employeeid
31	2	41890	1	4	2
28	3	58130	0	0	7
47	2	57620	9	9	14

Tabla 3: Renuncia de empleados con variables originales

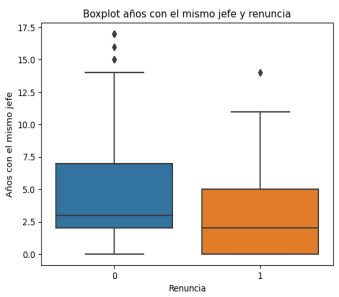
	employeeid	renuncia	0	1	2	3	4	5	6
1	2	1	1	0	0	0	0	0	0
6	7	1	1	1	1	1	0	0	0
13	14	1	1	0	0	0	0	0	0
28	29	1	1	1	0	0	0	0	0
30	31	1	1	0	0	0	0	0	0

Tabla 3: Renuncias con valores estandarizados

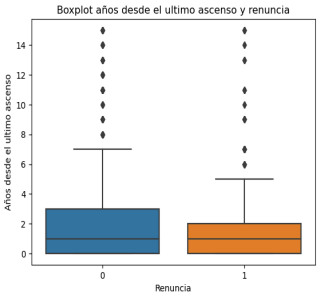
employeeid	renuncia	age	jobsatisfaction	monthlyincome	yearssincelastpromotion	yearswithcurrmanager
1	0	1,541561584	1,15613243	1,40628822	-0,681478414	-1,160491092
2	1	-0,652749629	-0,665398455	-0,491283052	-0,371592564	-0,040637764
3	0	-0,543034068	-0,665398455	2,726744448	-0,681478414	-0,320601096
4	0	0,115259296	1,15613243	0,387037156	1,487722537	0,239325568
5	0	-0,543034068	-1,576163897	-0,883891335	-0,681478414	-0,040637764

Tabla 4: ruta de decisión para cada empleado

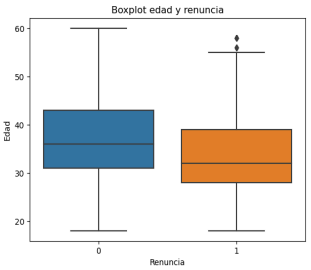
Según el algoritmo la variable más importante son los ingresos mensuales, observando la gráfica número 9 no se encuentran diferencias puntuales, esto puede ser causado por la sensibilidad de los árboles de decisión, el árbol podría estar tomando relaciones más complejas, tener interacciones con otras variables o tal vez se necesite exploración más profunda. En otras variables como la edad y la satisfacción la diferencia podría estar marcada de mejor forma. Por ejemplo para el empleado 7 se tiene una ruta en donde pasa por los nodos 0, 1, 2 y 3 inicialmente (son más 700 nodos), consideremos el nodo 1, este nodo hace referencia a la característica de la edad, su umbral corresponde al valor de -0.59, donde se determina hacia qué lado de la hoja va, observando en su valor normalizado su edad es de -0.65, al ser menor va hacia la izquierda del nodo, para agregar hasta este modo llegaron 768 muestras, finalmente de esta forma la ruta seguirá con el nodo número dos y así sucesivamente hasta llegar al nodo hoja el cual realiza la predicción.



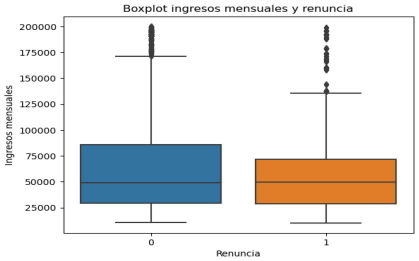
Gráfica 6: Años con el mismo jefe



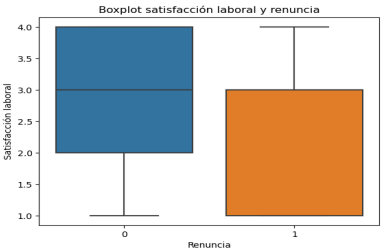
Gráfica 7: Años desde último ascenso



Gráfica 8: edad



Gráfica 9: Ingresos mensuales



Gráfica 10: Satisfacción laboral

Estrategias

Ingresos mensuales: evaluar la remuneración para los cargos específicos de las personas que predijo el modelo, estudiar si estos salarios son acordes a las ofertas presentes en el mercado, además ofrecer incentivos y beneficios por metas logradas. Un salario competitivo puede retener a los empleados. Por último tiene mucho sentido que el salario sea la variable más importante del modelo, generalmente una mala remuneración afecta la satisfacción laboral y estos puestos presentan una alta rotación.

Edad: diseñar planes para reducir el sesgo de edad entre los empleados, en la mayoría de casos las personas jóvenes ingresan a las empresas con un conocimiento nuevo e innovador y personas mayores bloquean u omiten esta información inhibiendo el desarrollo profesional de los empleados más jóvenes. Generar planes de acompañamiento a las personas cerca de jubilarse, generalmente estas personas se quedan en las empresas hasta obtener su pensión. Crear centros de dispersión y recreación de acuerdo a los grupos de edad,

Años con el mismo jefe: Diseñar capacitaciones en liderazgo, trabajo en equipo y manejo de personal para los jefes de las áreas o para empleados que tengan personal a cargo. Agregar a las encuestas la calificación a los jefes por parte de los empleados, a partir de esta evaluación tomar medidas de cambios de personal. incluir en las encuestas variables como acoso laboral, explotación laboral, discriminacion, inseguridad, falta de reconocimiento, entre otros.

años desde la última promoción: Generar políticas para el ascenso, establecer criterios objetivos para que todas las personas puedan participar de estas vacantes, involucrar a los empleados en proyectos que puedan aportar a su hoja de vida y aumentar su probabilidad de ser ascendidos.

Satisfacción laboral: incentivar el desarrollo profesional a través de proyectos y ascensos, reconocer y premiar los empleados, flexibilidad en los horarios de trabajo y la forma de trabaja (presencial, virtual, híbrido), ofrecer salarios justos, mejorar el

clima laboral por medio de salidas de campo, áreas de juegos recreativos e integraciones.

Finalmente con la ruta de decisión realizada para cada empleado se puede ser más precisos en la estrategia que se debería implementar para cada uno de ellos, sin embargo analizar más de 700 renuncias individualmente puede ser muy complejo para el área de recursos humanos.

CONCLUSIONES

La deserción laboral es un tema crítico para cualquier empresa y debe ser abordado desde herramientas analiticas que no solo puedan predecir con precisión la deserción si no también que puedan explicar el por qué de las decisiones de los empleados, siendo este el componente más importante para mitigar la problemática, además encontrar características y relaciones entre variables imperceptibles para un área de recursos humanos sin la aplicación de un componente analitico. Para el año 2017 estarán renunciando el 15,9% de los empleados, superando el valor anual del 15%, además se espera que con los planes de mitigación no supere el 12% la deserción en dicho año.

El área de recursos humanos ha sido una de las partes de las empresas beneficiadas de la nueva era de la inteligencia artificial, cada vez más se encuentran aplicaciones en este área, como por ejemplo proyectos de predicción de desempeño laboral, selección de talentos, análisis de sentimiento de los empleados, segmentación de los empleados, entre otros casos.

Los modelos probados inicialmente varían en su complejidad y para este caso se utilizó uno de los modelos más simples el decision tree classifier, que permite una mayor flexibilidad, explicabilidad y una menor capacidad de cómputo para su uso.

RECOMENDACIONES

La empresa debe mejorar la remuneración en aquellos cargos en los que se haya demostrado que se tiene una alta rotación de personal debido a que los empleados desertan por su baja remuneración. Esto apoyado de ir haciendo incentivos de manera dinámica que ayuden a tener un empleado atento y motivado en su labor.

La empresa como se propone en las estrategias, debe empezar a realizar una integración de edades, y buscar acompañantes que tengan experiencia en la misma para que apadrinen a los empleados jóvenes que están desertando. Si bien el talento joven no puede ser desperdiciado, también se propone contratar en esos cargos de alta deserción personas con alta experiencia laboral.

Se busca fortalecer el core de recursos humanos, personas capacitadas que tengan las habilidades para determinar unos buenos líderes en la empresa, que realmente realizan la labor que se tiene en el perfil de cargo. Enfatizando en que como líderes tengan unas altas capacidades en habilidades blandas.