EXPLORACIÓN DE DATOS

Después del análisis exploratorio hay 5 variables que se eliminarán, ya que no aportan nada al modelo. La variable conteo de empleados es constante, al igual que edad>18 y horas de trabajo. El rendimiento no entrega datos dicientes, por lo tanto también se decide descartar. Y el id también se elimina, puesto que su conteo no es de importancia para la ejecución de los modelos.

Las variables empresas anteriores, años de trabajo, satisfacción entorno, satisfacción laboral y conciliación familiar laboral contienen nulos. Para el área de analitica se determinó no eliminar datos nulos, sinó que se utiliza la mediana para las primeras primeras cuatro variables, dado que la mediana no es afectada por los datos atípicos. Para la variable años de trabajo, se utilizará la media.

Análisis relación entre variables numéricas

En la gráfica de correlación vista en el notebook análisis exploratorio se identificaron varias correlaciones altas entre algunas variables, la edad se correlacionó con años de trabajo con un valor mayor al 60%, además presentaba correlación baja con otras variables como empresas anteriores, años en la empresa, años desde ascenso y años con el jefe actual. En el análisis se puede concluir que los empleados a mayor años de edad mayor años de experiencia, lo que genera esta alta correlación, lo mismo sucede con las demás variables, las personas con mayor años de edad generalmente han estado en más empresa anteriormente, llevan más años en la empresa, tienen más años desde el último ascenso y llevan más tiempo con el jefe actual. Finalmente por estas razones se decide eliminar la variable edad.

La variable años en la empresa tiene una correlación mayor al 60% con las variables años de trabajo, años desde el último ascenso y años con el jefe actual. Esta correlación se puede determinar porque los empleados que llevan mayor tiempo en la empresa generalmente suman más años de experiencia, además la mayoría de empleados que han ingresado a la empresa no han tenido un ascenso y por lo tanto tienen el mismo jefe desde entonces. Por estas razones se decide eliminar la variable años en la empresa.

La variable años desde el último ascenso ya presentaba correlación con las variables analizadas al inicios, esta además tiene una correlación mayor al 40% con la variables años de trabajo y años jefe actual, esto se puede dar porque como se dijo anteriormente los empleados desde que obtuvieron su último ascenso tiene el mismo jefe desde entonces, además si se tienen más años en el puesto actual esto suma a los años de experiencia de la persona.

Finalmente las variables número de empresas anteriores, edad, años de trabajo, años en la empresa, años con el jefe actual, años desde el último ascenso pueden
estar explicando lo mismo de nuestra variable objetivo, están relacionadas positivamente, es decir que cuando aumenta una de estas variables la otra aumentará. Para evitar problemas de multicolinealidad en el modelo se determinó continuar sólo con las variables años con el jefe actual y los años de trabajo de los empleados.

Análisis relación entre variables categóricas

Se aplicó la prueba chi cuadrado para filtrar las variables entre las cuales el valor-p sea mayor que 0.05, se observó que la variable género tiene alta correlación con las variables deserción, departamento, formación y nombre de cargo. En el gráfico visto en el análisis exploratorio se puede observar cómo para cada departamento independiente de la cantidad de empleados es aproximadamente igual el número de hombres y mujeres. Por esta razón se decide eliminar la variable de género.

SELECCIÓN DE VARIABLES

Luego del análisis exploratorio realizado a la base de datos, en donde se aplicó una metodología de selección de variables preliminar llamada la selección univariante y se obtuvo una tabla general, ahora se aplicó un método integrado con la regresión Lasso para penalizar las variables que se va a llamar tabla 1 y un método wrapper basado en RFE que realiza una eliminación recursiva de características que se va a llamar tabla 2, finalmente obteniendo dos tablas bases para probar los modelos.

MODELOS

En el estudio por encontrar el mejor modelo que predijera la deserción de los empleados, se construyó inicialmente un modelo base de regresión logística con la base general de datos sin métodos de selección de variables y sin balanceo de clases, además se construyó el mismo modelo base pero con balanceo de clases con el hiperparámetro class weight. Segundo se construyeron tres modelos de regresión lineal, dos modelos de random forest classifier, dos modelos de gradient boosting classifier y por último dos modelos de support vector machine, obteniendo un total de 9 modelos, para cada tipo de modelo se le aplicó a la base 1 y la base 2, en el de regresión logística se probó un modelo de más con balanceo de clases manual. Finalmente se optimizaron los hiperparámetros de tres modelos, del random forest con base 1, el gradient boosting classifier con base 1 y el support vector machine con base 1.

RESULTADOS

	TRAIN			TEST		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Modelo base sin balanceo	0,862	-	-	0,851	0,719	0,262
Modelo base con balanceo	0,735	-	-	0,711	0,345	0,711
Regresión logistica Integrado	0,751	-	-	0,707	0,34	0,698
Regresión logistica Wrapper	0,741	-	-	0,703	0,338	0,711
Regresión logistica balanceo manual	0,773	-	-	0,705	0,335	0,679
Random Forest Integrado	0,808	0,44	0,77	0,781	0,43	0,73
Random Forest Integrado - Tuning hiperparámetros	0,793	0,4	0,65	0,758	0,39	0,62
Random Forest Wrapper	0,813	0,44	0,74	0,782	0,43	0,72
Gradient Boosting Classifier Integrado	0,909	0,92	0,89	0,836	0,54	0,54
Gradient Boosting Classifier Integrado - Tuning hiperparámetros	1	1	1	0,989	0,98	0,96
Gradient Boosting Classifier Wrapper	0,907	0,91	0,9	0,849	0,56	0,67
Support Vector Machine Integrado	0,749	0,36	0,78	0,714	0,35	0,71
Support Vector Machine Integrado - Tuning hiperparámetros	0,992	0,96	1	0,977	0,89	0,97
Support Vector Machine Wrapper	0,741	0,35	0,77	0,702	0,34	0,74

Tabla 1: Tabla de resultados de modelos

Nota: el modelo base se trata de una regresión logística con la tabla exportada luego del análisis exploratorio.

A continuación se harán varios análisis comparativos entre modelos.

Modelo base sin balanceo de clases vs modelo base con balanceo de clases

En esta primera comparación se deseaba encontrar las diferencias que se encontraban en un modelo de regresión logística al aplicar el hiperparámetro class weight, es decir al balancear las clases de la bases de datos, en los resultados se obtuvo que el accuracy en el modelo base sin balanceo en el test es 0,851 y en el modelo con balanceo de clases 0,711. Esto sucede porque el modelo es muy bueno prediciendo cuando no renuncia un empleado, ya que en los datos se está encontrando en su mayoría empleados que no renuncian, por lo tanto este modelo sin balancear no generaliza.

De aquí para adelante cuando se mencione el modelo base se trata del modelo de regresión logística con el balanceo de clases.

Modelo base sin selección de variables vs otros modelos con selección de variables

Recordemos que el accuracy significa la exactitud con la que el modelo clasifica correctamente, es decir las veces que acierta, además en esta comparación no se incluye el tuning de hiperparametros. Para esta comparación todos los modelos obtuvieron un mejor accuracy en el entrenamiento, sin embargo desmejoró en muy poca cantidad el accuracy en test para el caso de la regresión logística para ambas bases y el support vector machine con la base 2, en los demás se demostró que la selección de variables aporta significativamente a los resultados del modelo, esto se puede evidenciar en tabla superior.

Modelo de regresión logística con balanceo de clases de hiperparámetro vs modelo de regresión logística con datos de entrada balanceados

En esta comparación se quería poner a prueba cómo se podían obtener mejor resultados, a partir del hiperparámetro class weight o el balanceo manual con la función SMOTETomek(), esto se probó por medio del modelo de regresión logística, en la exactitud del modelo se observa que los dos en el test tienen aproximadamente un accuracy del 70%, por lo que se concluye que seguir trabajando con el hiperparametro o con los datos balanceados a partir del función seguirá cumpliendo de buena forma para obtener buenos resultados.

Método de selección de variables: integrados (Lasso) vs wrapper (RFE)

Para determinar qué método de selección de variables tendría mejor resultados se realizó un promedio para cada método de los 4 modelos en donde se aplicó (regresión logistica, random forest, gradient boosting y support vector machine), el promedio del accuracy del método integrado en los test es del 75.95% y en el método wrapper es 75.90%, concluyendo que los dos métodos obtienen una exactitud promedio muy similar, sin embargo esto podría variar dependiendo el modelo. Por las razones anteriores el tuning de hiperparametros se realizará a los modelos random forest, gradient boosting y support vector machine únicamente con la base 1, es decir con las variables predictoras elegidas por el método integrado.

Modelos sin tuning de hiperparametros vs modelos con tuning de hiperparametros

El método para buscar los mejores hiperparametros es la búsqueda en cuadrícula, se aplicó para los modelos random forest, gradient boosting y support machine, en todos la métrica de evaluación elegida para encontrar el mejor modelo fue el R2. Los R2 se exponen a continuación:

Random forest classifier: -3,4%
Gradient boosting classifier: 96,95%
support vector machine:63,99%

El gradient Boosting fue el modelo que mejor hiperparametros encontró obteniendo un R2 muy alto, es decir que este modelo es capaz de predecir en un 97% si un empleado renuncia o no. Luego el support vector machine fue el segundo modelo con mejores resultados, gracias también a la optimización de estos hiperparametros. Finalmente para el modelo random forest los hiperparametros desmejoraron el modelo sin tuning, se debe de buscar nuevos hiperparámetros para mejorar los resultados.

Análisis de overfitting

El modelo que tiene mayor sobreajuste es el gradiente boosting classifier, con la base 1 obtuvo un accuracy en train de 90,9% y en test 83,6%, para la base 2 obtuvo en train 90,7% y en test 84,9%, sin embargo la diferencia entre los valores del train y test no están tan alejados, por lo que se puede continuar con alguno de los dos modelos, sin embargo los resultados mostraron que este modelo con el tuning de hiperparametros aprendía y lograba generalizar.

Modelo seleccionado

Teniendo en cuenta que las métricas para seleccionar el modelo principalmente eran el accuracy y el recall se concluyó que el mejor modelo es el gradient boosting classifier, sus métricas fueron las siguientes:

Accuracy: la exactitud del modelo es de 98,9%, es decir que el modelo acierta el 99% de las veces, el modelo está generalizando.
Recall: el modelo es capaz de identificar el 96% de los empleados que van a renunciar.
precisión: un 98% de los empleados predichos que van a renunciar realmente lo harán, es decir que el modelo solo se equivocara en un 2% de las veces que un empleado va a renunciar.
f1-score: 99%, asume la importancia de la precisión y la exhaustividad del modelo.

Es importante resaltar que la importancia del recall se respalda en que para la empresa es vital poder predecir bien sobre las personas que en verdad van a renunciar, para poder tomar posturas anticipadamente y evitar la deserción, mejorando el área de recursos humanos, poder darle continuidad a proyectos de sus empleados antiguos y evitando tiempos muertos para sus nuevos empleados mientras se adaptan. A continuación se mostrará la matriz de confusión del modelo de gradient boosting con la base 1 pero sin tuning de hiperparametros (figura 1), además el mismo modelo pero con el tuning de hiperparametros (figura 3)

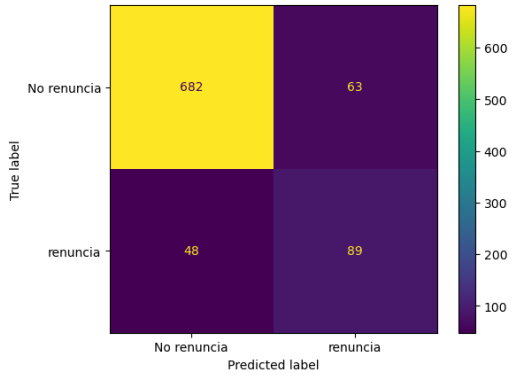

Figura 1: Matriz de confusión Gradient Boosting base 1 sin tuning

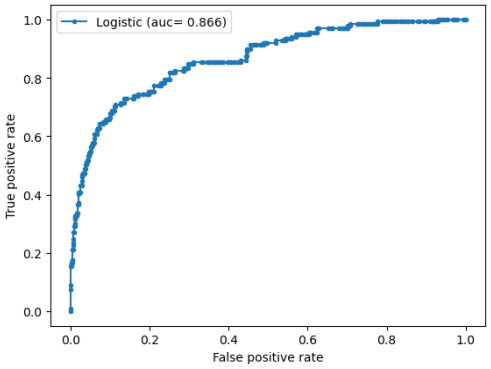

Figura 2: Gráfico de falsos positivos Vs falsos negativos Gradient Boosting base 1 sin tuning

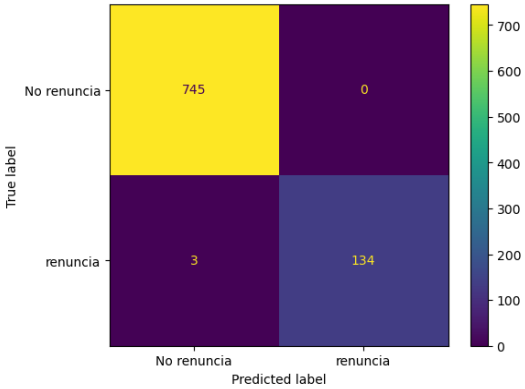

Figura 3: Matriz de confusión Gradient Boosting base 1 con tuning

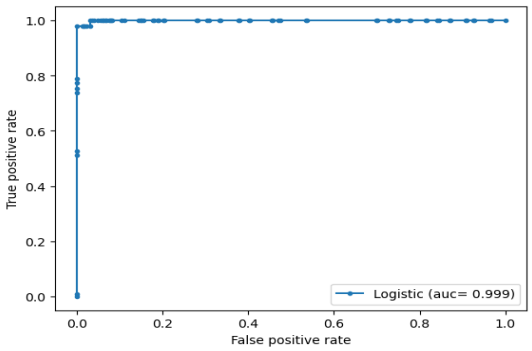

Figura 4: Gráfico de falsos positivos Vs falsos negativos Gradient Boosting base 1 con tuning

En la matriz de confusión se observa que los true positives y true negative cuando se hace tuning de hiperparametros aumentan, además el número de falsos positivos y falsos negativos disminuye drásticamente, evitando cometer el error tipo 1 y tipo 2.
En la curva ROC el área bajo la curva aumenta casi llegando a 1, la medida de sensibilidad pasa de 86% a un 99%, lo que indica que el modelo es capaz de distinguir entre las clases, es decir predecir correctamente la deserción de un empleado.

En general todos los modelos realizados, sin incluir el base, los modelos de gradient boosting y el modelo optimizado de support vector machine, obtienen mayor sensibilidad que precisión, es decir que los modelos cometen más el error tipo 1. En los modelos optimizados menos el de random forest se evitan en gran cantidad los escenarios de cometer el error tipo 1 y tipo 2, es decir que tienen un buenas métricas en precisión y sensibilidad, el modelo de random forest se debe hacer tuning nuevamente. En la tabla 1 se puede realizar estas comparaciones.

VARIABLE PREDICTORA MÁS IMPORTANTE

En el análisis de la variable predictora más importante se basó en la importancia dada por el modelo de gradient boosting classifier, en orden de mayor a menor importancia se tiene que:

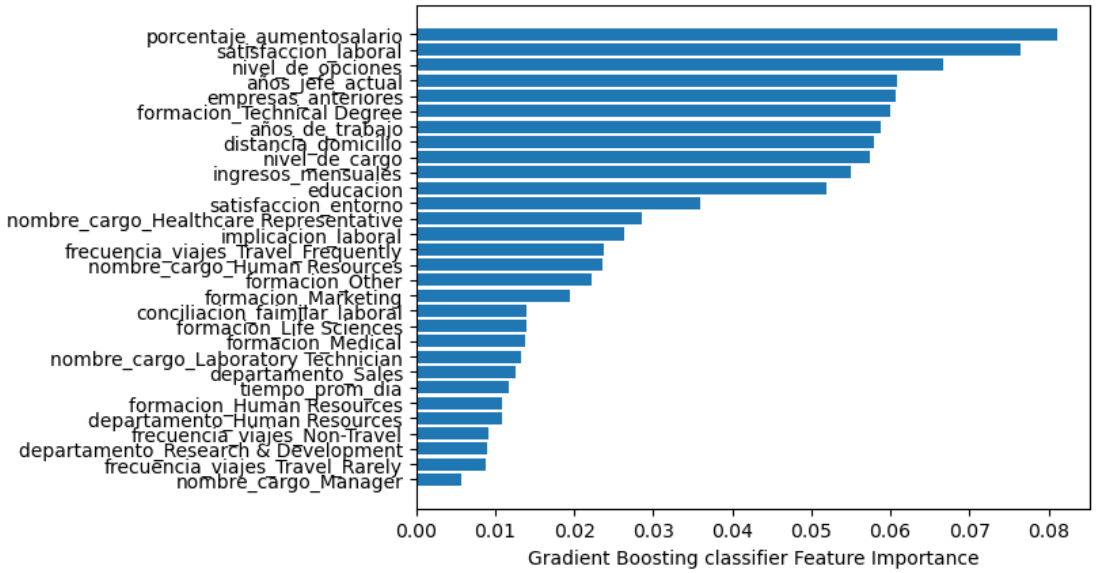

Figura 5: Gráfico de importancia Gradient Boosting

El modelo indica que la variable más importante en la deserción es el porcentaje de aumento salarial, además de segunda la satisfacción laboral.

CONCLUSIONES

En general todos los modelos pueden ser útiles para la empresa porque presentan buenas métricas de desempeño, algunos tienen mayor complejidad y por lo tanto necesitan mayor capacidad de cómputo, como lo es el gradient boosting classifier, sin embargo en la mayoría de casos obtienen mejores resultados que los algoritmos convencionales.

La empresa debe de utilizar el modelo de gradient boosting classifier para predecir oportunamente y correctamente si los empleados desertaran, ayudando a crear planes de mitigación de esta problemática desde el área de recursos humanos. Este modelo clasifica correctamente el 99% de las veces..

La variable más importante según el modelo de gradient boosting classifier es el porcentaje de aumento salarial, es decir que las personas que menor aumento tienen tienden a renunciar

fácilmente. La segunda variable más importante, es la satisfacción laboral lo que tiene mucho sentido, generalmente las personas aburridas en su trabajo tienden a renunciar más fácilmente y a rendir menos por lo que pueden ser despedidos.En la empresa se deben brindar aumentos salariales iguales para todos, además mejorar las condiciones del entorno de trabajo para mejorar la satisfacción de los empleados

RECOMENDACIONES

Se recomienda seguir haciendo nuevamente tuning de los hiperparametros del modelo de random forest, además realizarlo para el modelo regresión logística donde no se probó.

Los estimadores para los métodos de selección de variables pueden cambiarse para encontrar otras variables que aporten al modelo.

Se recomienda implementar modelos más avanzados como el Extreme Gradient Boosting