ANÁLISIS PREDICTIVO SOBRE LA DESERCIÓN DE EMPLEADOS

Elaborado por:

DIANA CATALINA VELÁSQUEZ GAVIRIA

JUAN CAMILO CEBALLOS ARIAS

SANTIAGO ARISTIZÁBAL TORO

Materia: Estadística Multivariada Avanzada

Profesor: TOMÁS OLARTE HERNÁNDEZ

Maestría en Ciencias de los Datos y Analítica

Universidad Eafit

Medellín 2020

PROYECTO DE APLICACIÓN ESTADÍSTICA MULTIVARIADA

- **Pregunta de investigación y objetivos**

La deserción de personal es un problema que enfrentan la mayoría de organizaciones por los costos que acarrea y los problemas de calidad y servicio que propicia, pues el personal nuevo que ingresa a una compañía no ha estado el tiempo suficiente para tener un desempeño óptimo. Por lo tanto, se hace necesario identificar los factores que causan la deserción y predecir la probabilidad de que ocurra para establecer acciones que la mitiguen.

- *Objetivo General*

Aplicar diferentes modelos de aprendizaje utilizando una base de datos tomada de la comunidad en línea de científicos de datos y profesionales del aprendizaje automático, Kaggle; en la que se tiene información sobre la rotación de personal en una empresa, con el fin de predecir la probabilidad de deserción de los mismos utilizando técnicas estudiadas en el curso de Estadística Multivariada y otras metodologías sugeridas en la literatura para el problema de Churn, y comprender conceptos como error real de entrenamiento, error de aprendizaje, regularización, entre otros.

- *Objetivos específicos*

- Realizar un análisis descriptivo de los datos para entender el comportamiento de sus variables previo al modelado.
- Realizar transformación de variables para el proceso de modelado.
- Realizar selección de variables para disminuir la complejidad de los modelos.
- Aplicar modelos de aprendizaje automático para la predicción de la deserción de empleados, empezando por los modelos más simples y siguiendo las recomendaciones vistas en clase.
- Balancear las clases de la variable respuesta mediante el aumento de la clase más pequeña y la disminución de la clase más grande, con el fin de modelar con nuevas muestras y hacer comparaciones.
- Escoger un modelo de todos los analizados de acuerdo con los mejores resultados en entrenamiento y test, luego de aplicar regularización para corregir el ajuste si es necesario.
- Obtener conclusiones sobre el análisis.

- **Revisión de la literatura, estado del arte y bibliografía**

Flores-Méndez et Al. (2018) abordan el problema de rotación de clientes de una empresa de comunicaciones, caracterizándolos con variables como edad, género,

lealtad, perfiles de uso, número de dispositivos, indicadores de satisfacción y tráfico de uso además de una red social mediante grafos en caso de que otras personas influyan en la decisión. Para el análisis utilizaron el modelado basado en agentes, una técnica que usa una población de agentes de comportamiento real interactuando entre sí. En este modelo se parametrizan las variables y luego se hace una simulación.
Usando datos de 2014 de la Oficina de Estadística de la Unión Europea, Eurostat, y calibrando los parámetros según su distribución inicial, realizaron una simulación para asignar los clientes a diferentes empresas de telecomunicaciones e iban cambiando con el tiempo. Cada variable fue calculada con base en otras variables, por lo que el modelo fue dividido en módulos.
Los hallazgos evidenciaron que, en un escenario con redes sociales, la información fluía más rápido, por lo que los agentes tendían a estar en la empresa de telecomunicaciones donde más gente cercana a ellos estuviera.

Por otro lado, Kim, M et Al. (2017) usan datos de clientes de una empresa de telecomunicaciones de Corea del Sur para predecir la rotación. La base de datos contiene información desde 2010 hasta 2014, periodo en el cual la mitad de las personas canceló el servicio.
Estimaron la variable uso de contenido con hipótesis y con un modelo de regresión curvilínea mediante variables como el paquete de servicios contratado, meses restantes de contrato, puntos de membresía, años de membresía y el historial de quejas. Luego estimaron la rotación de clientes con una hipótesis y un modelo logit usando algunas de las otras variables, junto con la cantidad de canales vista, y otros servicios contratados o consumidos relacionados con el internet.
Las variables resultantes fueron: el paquete de servicios, el pago por visitas con una relación negativa en la rotación y el historial de quejas, junto con el número de canales por tema comprados mensualmente con una relación positiva en la rotación. Las otras variables fueron irrelevantes o no significativas.

Finalmente, se indagaron algunas soluciones para el problema de Churn en Kaggle. Entre las técnicas con las cuáles se ha intentado abordar, están el bosque aleatorio, el Gradient Boosting, los K Vecinos Más Cercanos y las Máquinas de Soporte Vectorial, todos en su versión para clasificación.

- **Metodología de Investigación**

Para llevar a cabo el análisis se utilizará la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), que consiste en:

- ✓ Comprensión del negocio.
- ✓ Estudio y comprensión de los datos.
    - Composición de la base de datos (cantidad de variables, total de registros, tipos de variables).
    - Análisis descriptivo (variables numéricas y categóricas).
    - Selección de variables.

- ✓ Modelado.
  - Implementación de modelos
  - Balanceo de datos
  - Comparación de modelos
- ✓ Evaluación de resultados.
- ✓ Despliegue.

- **Análisis de los datos**

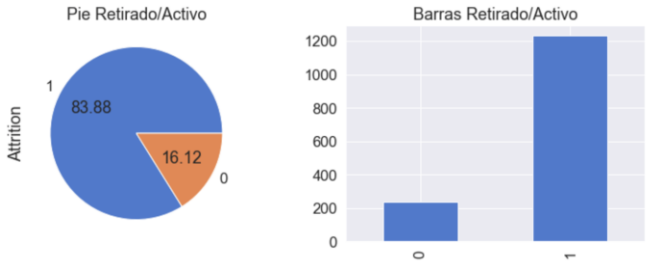
La base de datos utilizada para llevar a cabo el ejercicio es tomada de Kaggle y consiste en una base ficticia creada por los científicos de IBM. Este dataset consta de 1470 registros y 31 variables, entre las cuales se encuentra una con información referente a si el empleado dejó o no la compañía y las demás características hacen referencia a: rol del empleado en la compañía, información sobre la satisfacción del empleado, información relacionada con el salario mensual, diario, etc. Cambios salariales de un año con respecto al siguiente, información de desempeño y frecuencia de viajes laborales.

Como primer paso, se intentó entender los datos de manera general, para ello se realizó un análisis descriptivo: medidas de tendencia central, proporción de la variable respuesta (Attrition), análisis univariado de las variables numéricas, análisis de las variables numéricas agrupando por clases de la variable respuesta, análisis de las variables categóricas agrupando por clases de la variable respuesta, histogramas, boxplots, moisac plots, matriz de correlaciones y transformación de variables categóricas a dummies, como se muestra en la siguiente sección.

- **Estudio y comprensión de los datos**

Se define que la variable de interés o variable respuesta es "*Attrition*", que toma los valores de 0 para empleados retirados y 1 para empleados activos y está conformada así:

Figura 1. Visualización variable respuesta: "Attrition".



Fuente: creación propia

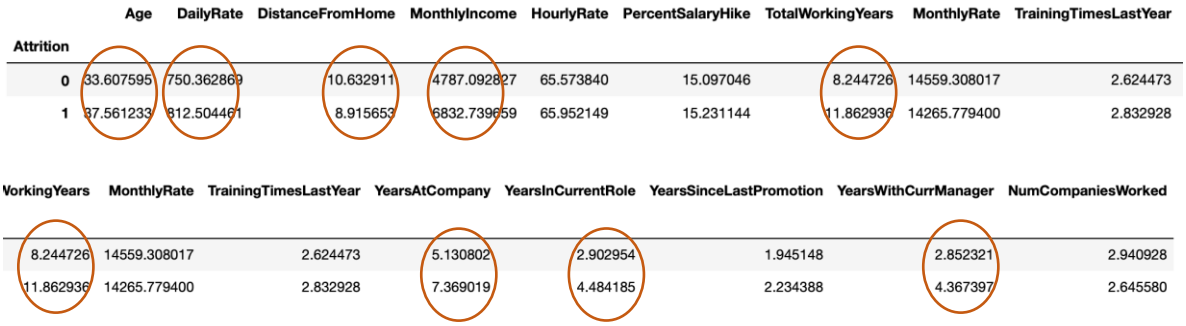
El 16.12% de los empleados ha dejado la compañía, lo cual corresponde a 237 empleados y el 83.88% permanece en ella, equivalente a 1233 empleados.

Posteriormente se realiza un análisis de cada una de las variables de la base con respecto a la variable de interés, dividiendo el conjunto de datos original en dos subsets, uno para variables numéricas (14 variables) y otro para variables categóricas (16 variables):

- *Análisis de variables numéricas*

En el conjunto de variables numéricas se analiza el promedio de cada variable para las dos clases de *Attrition* así:

Figura 2. Promedio de cada variable para las clases de “Attrition”.



Attrition	Age	DailyRate	DistanceFromHome	MonthlyIncome	HourlyRate	PercentSalaryHike	TotalWorkingYears	MonthlyRate	TrainingTimesLastYear
0	33.607595	750.362869	10.632911	4787.092827	65.573840	15.097046	8.244726	14559.308017	2.624473
1	37.561233	812.504461	8.915653	6832.739659	65.952149	15.231144	11.862936	14265.779400	2.832928

VorkingYears	MonthlyRate	TrainingTimesLastYear	YearsAtCompany	YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager	NumCompaniesWorked
8.244726	14559.308017	2.624473	5.130802	2.902954	1.945148	2.852321	2.940928
11.862936	14265.779400	2.832928	7.369019	4.484185	2.234388	4.367397	2.645580

Fuente: creación propia

Con este análisis se busca entender cómo varía el promedio tanto para empleados retirados como para los activos e identificar las variables donde se presenta la mayor diferencia a simple vista; en la Figura 2, representadas con un círculo naranja.

Además, para complementar el entendimiento se realizan boxplots comparativos que permiten observar la distribución y variabilidad de los datos en cada categoría de *Attrition*, encontrando que pareciera haber diferencia entre ambas y se da una idea inicial sobre qué variables pueden ser de interés para explicar la deserción de los empleados.

Continuando con el análisis numérico, se realiza un histograma de dichas características de forma univariada para comprender sus distribuciones y se encuentra que algunas variables como DailyRate, HourlyRate, MontlyRate presentan distribuciones muy uniformes.

Finalmente se realiza una matriz de correlación para estas variables y se encuentra que al parecer no existen problemas de multicolinealidad ya que, pese a que algunas variables presentan correlaciones, no se encuentran tan cercanas a 1.

- *Análisis de variables categóricas*

De la misma manera, se realiza un análisis sobre las variables categóricas con respecto a la variable respuesta para entender cómo varían las proporciones de cada clase entre los niveles de *Attrition*. Para llevarlo a cabo, se construyen mosaic plots y tablas de contingencia para identificar aquellas características donde dichas proporciones parecen diferentes en empleados activos y retirados, encontrando que las variables BusinessTravel, Department, EducationField, EnvironmentSatisfaction, JobInvolvement, JobLevel, JobRole, JobSatisfaction, MaritalStatus, OverTime, StockOptionLevel y WorkLifeBalance, parecen tener diferencias.

- *Modelado*

Posteriormente, se hizo la partición de los datos en los subconjuntos para entrenamiento, test y validación, con una proporción de 60%, 20% y 20% respectivamente. Luego, se hizo la estandarización de las variables para comenzar con el modelado.

Se utilizaron varios modelos de aprendizaje supervisado para comparar sus resultados, pero antes fue necesario seleccionar las variables predictoras relevantes para el estudio mediante dos técnicas: *Regularización Lasso*, cuya fundamentación fue estudiada en el curso y otra técnica de Machine Learning llamada *Eliminación Recursiva de Características con Validación Cruzada - Recursive Feature Elimination with Cross-Validation (RFECV)*: a partir de un modelo externo que asigna pesos a las variables, por ejemplo, los coeficientes de un modelo lineal o las importancias derivadas de un árbol de decisión, RFECV consiste en seleccionar variables considerando recursivamente conjuntos de características más y más pequeños. Primero, el estimador es entrenado con un conjunto inicial de variables y la importancia de cada una es obtenida. Después, las características menos importantes son eliminadas del conjunto que está siendo utilizado. Este procedimiento se repite de manera recursiva sobre el conjunto de variables eliminadas hasta que se alcanza el número deseado. La eliminación se hace en un bucle de validación cruzada para encontrar el número óptimo de características (Scikit learn, 2019).

Las dos bases de datos resultantes de cada uno de los métodos fueron usadas para evaluar los siguientes modelos: Regresión Logística, Gradient Boosting, Árbol de Decisión, Máquina de Soporte Vectorial, Bosque Aleatorio, K Vecinos Más Cercanos, un Proceso Gaussiano de Clasificación y una Red Neuronal Feedforward Clasificadora. Entre otras métricas, se evaluó el roc_auc en todos para hacer comparaciones.

La fundamentación conceptual de los modelos probados fue estudiada en el curso de Estadística Multivariada excepto la del Proceso Gaussiano y, por lo tanto, se explica a continuación:

Un Proceso Gaussiano es una distribución de probabilidad sobre variables aleatorias que cumplen que cualquier subconjunto finito de ellas tiene una distribución normal conjunta. Los parámetros del proceso son una función media m(x) y una matriz de covarianza k(x,x'), que a su vez es una matriz Kernel de Gram. Los Procesos Gaussianos tienen la siguiente forma: dadas las variables $x_1, \dots, x_N$, entonces el vector $f(x) \sim GP(m(x), k(x, x'))$. $f = (f(x_1), \dots, f(x_N))^T \sim N_N(0, K)$, donde K es la matriz de covarianza.

Para realizar clasificación con Procesos Gaussianos, se parte de un conjunto de entrenamiento $\{(x_i, y_i) | i = 1, \dots, n\}$ con $y_i = y(x_i)\ \epsilon\ \{-1,1\}$, siendo {-1,1} las clases y $x*$ y $y*$ datos de prueba. Para clasificar correctamente, buscamos estimar la función $\pi(x*) = p(y* = 1|x*)$.

El Proceso Gaussiano de clasificación $f(x) \sim GP(m(x), k(x, x'))$, define una función aleatoria f que tiene valores reales, pero debido a que las probabilidades oscilan entre 0 y 1, entonces se necesita una función σ: $(-\infty, \infty) \rightarrow [0,1]$ suave y creciente. Esta función será la logística, cuya fórmula es la función sigmoide:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Así, se puede escribir $\pi(x*) = p(y* = 1|x*) = \sigma(f(x*)$

El Proceso Gaussiano de clasificación se divide en dos pasos y usa la metodología de selección bayesiana de modelos.

El primer paso es estimar la distribución posterior del Proceso Gaussiano

$p(f*|X, y, x*)$ y después la probabilidad en la que se está interesado,

$\pi(x*) = p(y* = 1|x*) = \sigma(f(x*)$:

$$p(f*|X, y, x*) = \int p(f*|X, y, x*)\ p(f|X, y)\ df \quad (1)$$

$$\bar{\pi}* = p(y* = 1|X, y, x*) = \int \sigma(f*)\ p(f*|X, y, x*)\ df* \quad (2)$$

En la ecuación (1), la parte derecha tiene una integral que no distribuye normal y la ecuación (2) no se puede evaluar en general, por lo que no se puede hallar una forma analítica para la distribución para ninguna de las ecuaciones, por lo que es necesario aproximar $p(f|X, y)$ con una distribución normal para poder evaluar las integrales de forma aproximada. Esta aproximación, se realiza mediante la

aproximación de Laplace y otros métodos. Finalmente, se obtienen las estimaciones de las ecuaciones para realizar el cálculo.

Los resultados son los siguientes:

$$E[f * |X, y, x *] = k(x *)^T \nabla \log p\left(y\middle|\hat{f}\right)$$

$$V_q[f * |X, y, x *] = k(x *, x *) - k(x *)^T (K + W^{-1})^{-1} k(x *)$$

Donde $\hat{f}$ es el vector convergente de la aproximación de Laplace, $\nabla \log p\left(y\middle|\hat{f}\right)$ es el gradiente de la función $\log p\left(y\middle|\hat{f}\right)$, $W = -(\nabla\nabla \log p(f|X, y + K^{-1})$ y $\nabla\nabla \log p(f|X, y)$ el Hessiano de la función $\log p(y|X, Y)$

Con las anteriores aproximaciones, se puede calcular un valor cercano a las integrales de las ecuaciones (1) y (2), resolviendo el problema de clasificación binaria con Procesos Gaussianos (Velasco, 2017).

- *Balanceo de clases*

Posteriormente, debido a la diferencia en proporciones de las clases de la variable respuesta (83.88% contra un 16.12%) y con el fin de hacer más comparaciones, se hizo balanceo de clases de dos maneras: aumento de la minoría (oversampling) y disminución de la mayoría (undersampling). También se intentó modelar utilizando el hiperparámetro *class_weights* para que el modelo ajustara automáticamente el peso de las clases inversamente proporcional a su frecuencia, pero los resultados no presentaron alguna mejora. Se seleccionaron las características utilizando solamente la técnica RFECV y, otra vez se evaluaron los modelos de aprendizaje supervisado utilizados anteriormente.

La tabla 1 muestra los resultados obtenidos teniendo en cuenta la métrica ROC_AUC. Los valores resaltados representan el mejor resultado para cada uno de los tres casos.

Tabla 1. Métrica ROC_AUC con su respectivo balanceo de clases

Modelo	**ROC_AUC sin balanceo de clases**	**ROC_AUC con aumento de minoría**	**ROC_AUC con disminución de mayoría**
Regresión logística	**0.692**	0.771	**0.754**
Potenciación del gradiente	0.601	**0.966**	**0.748**
Árbol de decisión	0.575	0.912	0.631
Máquina de soporte vectorial	**0.693**	0.914	**0.754**
Bosque aleatorio	0.562	**0.966**	0.733
K vecinos más cercanos	0.547	0.813	0.678

Proceso Gaussiano de clasificación	0.555	0.921	0.681
MLP	**0.69**	0.935	0.72

Fuente: cálculos propios

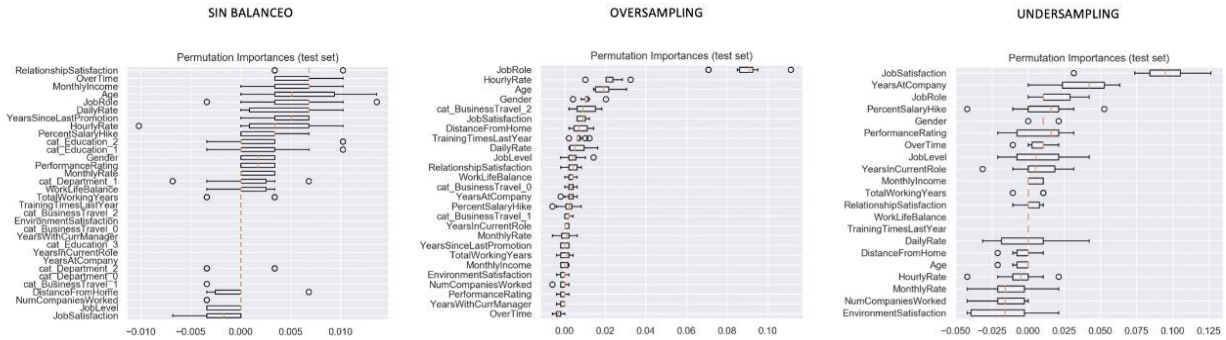
Los resultados obtenidos con balanceo de clases fueron mejores que sin aplicar esta técnica. Los resultados con aumento de la minoría fueron mejores que aquellos con disminución de la mayoría.

Debido al buen resultado en el modelado con balanceo de clases aumentando la minoría, se aplicaron las técnicas *Random Forest Importance* y *Permutation Importance* para entender mejor cuáles características fueron las más importantes en cada ensayo (con o sin balanceo) y qué sentido tienen dentro del contexto del problema analizado.

Random Forest Importance proporciona la importancia que le da el modelo a las variables, haciendo el cálculo con base en la impureza en problemas de clasificación; sin embargo, este presenta una debilidad pues tiene un sesgo hacia las variables numéricas (dándole más importancia con respecto a las variables binarias o categóricas de pocas clases) y puede darle mayor importancia a variables que no son predictoras. Por su parte, Permutation Importance corrige estos problemas ya que calcula la importancia de las variables a partir del decrecimiento en el performance del modelo cuando los valores de una característica son mezclados aleatoriamente (Scikit learn, 2019).

Teniendo en cuenta las limitaciones de Random Forest Importance, se decidió analizar la importancia de las variables con la técnica Permutation Importance. La figura 3 muestra los resultados para cada caso de balanceo de clases.

Figura 3. Permutation Importance para cada balanceo de clases.



Fuente: creación propia

Las variables más importantes resultaron ser el cargo en el trabajo, el horario, el género, el salario, la distancia al hogar y los años en la empresa. A pesar de que el modelo con oversampling presentaba un incremento significativo en las métricas de desempeño, causando dudas o sospechas sobre la veracidad de esta mejora, al analizar el Permutation Importance se evidencia que el score de importancia presenta mayor estabilidad para las diferentes repeticiones y además, las variables que resultan más relevantes tienen sentido en la vida laboral de una persona real, ya que pueden influir en permanecer o no en una empresa por lo que se concluye que los modelos con oversampling sí tuvieron el mejor rendimiento.

El modelo más estable fue la Regresión Logística, cuyo valor de AUC fue parecido al de la Máquina de Soporte Vectorial y al de la Red Neuronal con una capa y una neurona. Los modelos que tuvieron mejor desempeño fueron el Gradient Boosting con aumento de la minoría y el Random forest.

Cabe mencionar que también se calcularon otras métricas como Presición, Recall y F1 score, las cuáles se pueden encontrar en los notebooks adjuntos.

- **Uso de la metodología y herramientas de aprendizaje estadístico**

El trabajo fue realizado con el lenguaje de programación Python, mediante el ambiente Jupyter Notebok. Se usó la biblioteca Pandas para importar los datos iniciales, además de ejecutar ciertas labores en el análisis descriptivo. También se importaron las bibliotecas Matplotlib y Seaborn para graficar variables. La separación y estándarización de los datos, el modelado y la obtención de métricas se realizó con distintas funciones de la biblioteca Scikit-learn, muy útil para aprendizaje automático. Se escogieron tales bibliotecas por su facilidad de uso.

Luego de la estandarización de los datos, se procedió con la selección de características (feature selection), pues inicialmente se contaba con 30 variables explicativas y tal número podría causar problemas de multicolinealidad o de complejidad dimensional. Este proceso de realizó utilizando Lasso y RFECV como ya fue mencionado.

Los modelos fueron seleccionados según el procedimiento recomendado para Aprendizaje de Máquina, esto es, empezar con la evaluación del modelo más sencillo posible (regresión logística para el caso de una clasificación binaria) y posteriormente usar modelos más complejos para hacer comparaciones. Al tratarse de un problema de clasificación, se probó el modelo de potenciación del gradiente, algo más complejo que la regresión logística por su función de pérdida y su optimización. Se siguió con un árbol de decisión, modelo con un enfoque distinto a

los anteriores, pero relativamente simple. Después se evaluó la máquina de soporte vectorial, un modelo más complejo y que, antes del balanceo de clases, pese a tener un buen resultado no había sido mejor que la regresión logística inicial, por lo que se trató con otros modelos para comprobar si mejoraban con respecto a dicha regresión. Así, se probaron los K vecinos más cercanos y el proceso gaussiano de clasificación, pero finalmente sin balanceo de clases, ninguno de los modelos fue mejor que la regresión logística.

Posteriormente se calcularon otras métricas para su comparación, pero siempre teniendo como referencia la métrica ROC_AUC, la cual se seleccionó porque en problemas de clasificación binaria es muy útil para medir qué tan bien puede el modelo separar las clases.

El balanceo de clases se realizó para estudiar cuál era el comportamiento de los mismos modelos al ser implementados en muestras con proporciones iguales en las clases de la variable a predecir; inicialmente se intuía una mejora en el desempeño, situación que se quiso comprobar. El proceso de balanceo se ejecutó utilizando la biblioteca imblearn, de gran ayuda para aprendizaje con clases desbalanceadas, imbalanced learn.

La selección de características, feature selection, se realizó mediante regularización Lasso y RFECV como se mencionó anteriormente.

- **Entregables y su descripción:**

A parte de este informe, los demás entregables se encuentran en el GITHUB del proyecto y las instrucciones para leer los archivos se encuentran en el REAMDE del repositorio.

- **Conclusiones y trabajo futuro:**

Finalmente se observa que luego de realizar el pre-procesamiento, análisis de los datos, selección de variables y comparación de modelos, los mejores resultados se presentaron con el balanceo de clases con oversampling, específicamente en los modelos de Gradient Boosting y Random Forest, en los cuáles se observan valores similares tanto en el conjunto de test y validación en las distintas métricas y que no que no evidenciaron problemas de sobreajuste y varianza, habilitándolos para su implementación.

Para resolver problemas similares en el futuro, es importante tener en cuenta los adelantos que se tengan, por lo que se debe consultar constantemente la literatura para actualizarse sobre el estado del arte del objeto de estudio. Se pueden explorar varias técnicas de Feature Selection y usar modelos con todas las variables seleccionadas para explorar más opciones y comparar sus resultados. En caso de que las clases estén desbalanceadas, se puede intentar con otras técnicas de balanceo como SMOTE.

Por otro lado, un posible trabajo futuro podría ser la aplicación del problema de rotación de personal a uno de rotación de clientes. Para ello, podría clasificarse si un cliente sigue o no con la compañía, con el fin de mitigar su rotación y ayudar a conservar los ingresos corporativos. Otro posible trabajo para el futuro sería implementar modelos para rotación de productos, en caso de una empresa de retail o rotación de estudiantes para el caso de colegios y universidades.

En cuanto a la empresa, se deben implementar los modelos realizados. Es importante estudiar modelos de rotación de personal para evitar costos de indemnización, capacitación, disminución en la productividad y afectación en la atención a clientes y usuarios finales.

- **Ejecución del plan**

Tabla 2. Plan de ejecución original

			Marzo				Abril			
Tarea	Inicio	Final	sem1	sem2	sem3	sem4	sem1	sem2	sem3	sem4
✔ Comprensión del negocio.										
✔ Estudio y comprensión de los datos.										
✔ Composición de la base de datos (cantidad de variables, total registros, tipos de variables)										
✔ Análisis descriptivo (variables numéricas y categóricas)										
✔ Selección de variables										
✔ Modelado										
✔ Evaluación de resultados										
Despliegue										

Fuente: creación propia

Tabla 3. Plan de ejecución real

			Marzo				Abril				Mayo	
Tarea	Inicio	Final	sem1	sem2	sem3	sem4	sem1	sem2	sem3	sem4	Sem1	Sem2
✔ Comprensión del negocio.												
✔ Estudio y comprensión de los datos.												
✔ Composición de la base de datos (cantidad de variables, total registros, tipos de variables)												

✔ Análisis descriptivo (variables numéricas y categóricas)												
✔ Selección de variables												
✔ Modelado												
✔ Evaluación de resultados												
Despliegue												

Fuente: creación propia

En términos generales el cumplimento de tiempos del cronograma planeado se cumplió. Se pudo haber tenido en cuenta un poco más de tiempo para seleccionar las variables y para realizar el modelado. En general, la mayor parte del tiempo fue dedicada al pre-procesamiento de la información. En este caso los datos eran públicos, por lo que no hubo problemas de recolección. La ejecución de modelos normalmente es rápida, aunque a veces puede haber ciertos imprevistos. El mayor de ellos fue el despliegue, pues hubo que reunir toda la información y documentar el proceso.

- **Implicaciones éticas**

El mayor problema que puede afrontar esta clase de estudios está relacionado con la información del empleado. No es claro qué tan ético es que una empresa tenga acceso a cierta información personal del mismo. No es claro si es ético si a la hora de contratar un candidato con características personales con un riesgo alto de deserción laboral como vivir lejos del trabajo o estar casado, deba ser motivo para descartarlo. Por ejemplo, si determinado género o grupo de edad resultan tener una mayor probabilidad de rotación, ¿es ético descartar al candidato porque pertenece al grupo de alta probabilidad, aunque aparentemente sea bueno y cumpla con los requisitos para trabajar? Las empresas deben analizarlo y tomar la decisión.

El modelo aún no fue implementado en ninguna empresa, por lo que las verdaderas implicaciones éticas relacionadas con las decisiones a tomar en la compañía con el modelo implementado aún no son muy conocidas. Sin embargo, la recolección de cierta información sobre los trabajadores que es necesaria para emplear en el modelo no es una tarea fácil; muchos empleados son reacios a brindar sus datos a la compañía.

- **Aspectos legales y comerciales**

Es posible que estos modelos se limiten en alguna información personal dependiendo del marco jurídico del país o la región. Por ejemplo, en algunos lugares puede ser ilegal saber dónde vive el empleado, su estado civil o el número de hijos que tiene, lo que podría limitar un estudio de este tipo. En cuanto a la exposición de

los resultados, las compañías que implementan esta clase de modelos deben ser cautelosas, pues intervienen en el recurso humano, un tema sensible, pues existe la posibilidad de que, si un empleado se entera de que es catalogado como alta probabilidad de rotación, es posible que su rendimiento cambie, que deserte sin haberlo pensado o incluso en algunos casos podría demandar a la compañía. El potencial del modelo radica en mantener empleados estables a lo largo del tiempo, esto como se mencionó anteriormente, puede aumentar la productividad y rentabilidad de la empresa.

- **Bibliografía**

- Employee attrition via Ensemble tree-based methods. (s. f.). Recuperado 27 de abril de 2020, de https://www.kaggle.com/arthurtok/employee-attrition-via-ensemble-tree-based-methods
- Employee attrition modelling. (s. f.). Recuperado 27 de abril de 2020, de https://www.kaggle.com/sambitd/employee-attrition-modelling
- FinalProject. (s. f.). Recuperado 27 de abril de 2020, de https://www.kaggle.com/nidhishjain/predicting-job-termination
- Flores-Méndez, M. R., Postigo-Boix, M., Melús-Moreno, J. L., & Stiller, B. (2018). A model for the mobile market based on customers profile to analyze the churning process. Wireless Networks, 24(2), 409-422.
- Kaggle. (2018). Data files used for models (Versión 2) [Archivo de datos]. Recuperado de https://www.kaggle.com/carmelgafa/data-files-used-for-models
- Kim, M. J., Kim, J., & Park, S. Y. (2017). Understanding IPTV churning behaviors: focus on users in South Korea. Asia Pacific Journal of Innovation and Entrepreneurship.
- Predict employment termination. (s. f.). Recuperado 27 de abril de 2020, de https://www.kaggle.com/dredlaw/predict-employment-termination
- Scikit learn. (2019). Recursive Feature Elimination. Recuperado 1 de mayo de 2020, de https://scikit-learn.org/stable/modules/feature_selection.html#rfe
- Scikit learn. (2019). Permutation feature importance. Recuperado 23 de mayo de 2020, de https://scikit-learn.org/stable/modules/permutation_importance.html#permutation-importance
- Velasco Pardo, V. (2017). Aprendizaje multi-tarea mediante procesos gaussianos para clasificación (Bachelor's thesis).