**ANÁLISIS PREDICTIVO SOBRE LA DESERCIÓN DE EMPLEADOS**

Elaborado por:

DIANA CATALINA VELÁSQUEZ GAVIRIA

JUAN CAMILO CEBALLOS ARIAS

SANTIAGO ARISTIZABAL TORO

Materia: Estadística Multivariada Avanzada

Profesor: TOMÁS OLARTE HERNÁNDEZ

Maestría en ciencia de datos y analítica

Universidad Eafit

Medellín 2020

**PROYECTO**

* **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 deserción de los empleados 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 empleados 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 el problema 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 de 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. Transformación de variables categórticas a dummies. (Todo esto entregado en el anteproyecto)

En segundo lugar, 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 aplicaron varios modelos de aprendizaje supervisado para comparar sus resultados, pero antes fue necesario seleccionar las variables predictoras relevantes para el estudio. Se utilizaron dos técnicas: Regularización Lasso y Eliminación Recursiva de Características con Validación Cruzada (RFECV).

Las dos bases de datos resultantes de cada uno de los métodos fueron usadas para finalmente evaluar los siguientes modelos: regresión logística, gradient boosting, árbol de decisión, máquina de soporte vectorial, bosque aleatorio, K vecinos más cercanos y un proceso gaussiano de clasificación. Entre otras métricas, se evaluó el roc\_auc.

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 (oversapling) y disminución de la mayoría (undersampling). Esta vez, se seleccionaron las características utilizando solamente la técnica RFECV y, nuevamente se evaluaron los mismos 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 por columna.

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 | **0.835** | 0.855 | **0.845** |
| Potenciación del gradiente | 0.81 | 0.979 | 0.815 |
| Árbol de decisión | 0.718 | 0.876 | 0.77 |
| Máquina de soporte vectorial | 0.83 | 0.934 | 0.838 |
| Bosque aleatorio | 0.816 | **0.982** | 0.813 |
| K vecinos más cercanos | 0.736 | 0.857 | 0.811 |
| Proceso gaussiano de clasificación | 0.595 | 0.966 | 0.727 |

Fuente: cálculos propios

Los resultados obtenidos con balanceo de clases fueron mejores que sin aplicar esta técnica, excepto por el bosque aleatorio con disminución de la clase grande. Los resultados con aumento de minoría fueron mejores que aquellos con disminución de mayoría, entonces para el problema estudiado, resulta mejor trabajar sólo con este tipo de balanceo de clases.

El modelo más estable fue la regresión logística, que puede ser explicado porque los modelos lineales tienden a tener una varianza menor, aunque tengan una transformación.

El modelo que tuvo mejor desempeño fue el bosque aleatorio con aumento de minoría. Dicho modelo tiene, comúnmente, un buen desempeño en este tipo de aplicaciones pues al trabajar con muchos árboles y calcular un promedio de éstos, se tiene buena generalización, aunque puede tener ciertos problemas de varianza.

Cabe mencionar que también se calcularon otras métricas, pero siempre se tuvo como referencia el ROC\_AUC.

Las tablas 2,3 y 4 muestran, respectivamente, los resultados de las otras métricas sin balanceo de clases, aquellos con aumento de minoría y los obtenidos con disminución de mayoría.

Tabla 2. Resultados de otras métricas sin aplicar balanceo de clases

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Sin balanceo de clases** | | | |
| **Modelo** | **AUC** | **Precisión** | **Recall** | **F1 score** |
| Regresión logística | 0.635 | 0.755 | 0.635 | 0.667 |
| Potenciación del gradiente | 0.641 | 0.807 | 0.641 | 0.68 |
| Árbol de decisión | 0.608 | 0.656 | 0.608 | 0.624 |
| Máquina de soporte vectorial | 0.66 | 0.789 | 0.66 | 0.697 |
| Bosque aleatorio | 0.523 | 0.93 | 0.523 | 0.507 |
| K vecinos más cercanos | 0.535 | 0.931 | 0.535 | 0.528 |
| Proceso gaussiano de clasificación | 0.611 | 0.639 | 0.611 | 0.622 |

Fuente: cálculos propios

Tabla 3. Resultados de otras métricas aplicando balanceo de clases aumentando la minoría

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Aumento de minoría** | | | |
| **Modelo** | **AUC** | **Precisión** | **Recall** | **F1 score** |
| Regresión logística | 0.768 | 0.771 | 0.768 | 0.767 |
| Potenciación del gradiente | 0.948 | 0.948 | 0.948 | 0.947 |
| Árbol de decisión | 0.809 | 0.813 | 0.809 | 0.807 |
| Máquina de soporte vectorial | 0.887 | 0.887 | 0.887 | 0.887 |
| Bosque aleatorio | 0.956 | 0.956 | 0.956 | 0.955 |
| K vecinos más cercanos | 0.802 | 0.805 | 0.802 | 0.801 |
| Proceso gaussiano de clasificación | 0.912 | 0.916 | 0.912 | 0.911 |

Fuente: cálculos propios

Tabla 4. Resultados de otras métricas aplicando balanceo de clases disminuyendo la mayoría

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Disminución de mayoría** | | | |
| **Modelo** | **AUC** | **Precisión** | **Recall** | **F1 score** |
| Regresión logística | 0.736 | 0.736 | 0.736 | 0.736 |
| Potenciación del gradiente | 0.697 | 0.696 | 0.697 | 0.695 |
| Árbol de decisión | 0.699 | 0.699 | 0.699 | 0.695 |
| Máquina de soporte vectorial | 0.756 | 0.757 | 0.756 | 0.756 |
| Bosque aleatorio | 0.739 | 0.738 | 0.739 | 0.737 |
| K vecinos más cercanos | 0.72 | 0.72 | 0.72 | 0.716 |
| Proceso gausiano de clasificación | 0.694 | 0.693 | 0.694 | 0.694 |

Fuente: cálculos propios

* **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 de máquinas. 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 en aprendizaje automático, esto es, empezar con la evaluación del modelo más sencillo posible (regresión logística para el caso de una clasificación binaria) y posteriormente se trató con modelos más complejos y se hizo 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 automático con clases desbalanceadas, imbalanced learn.

La selección de características, feature selection, se realizó mediante regularización Lasso, cuya fundamentación fue estudiada en el curso de Estadística Multivariada 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 características considerando recursivamente conjuntos de características más y más pequeños. Primero, el estimador es entrenado con un conjunto inicial de características y la importancia de cada una es obtenida. Después, las características menos importantes son eliminadas del conjunto de variables que está siendo utilizado. Este procedimiento se repite de manera recursiva sobre el conjunto de variables eliminadas hasta que se alcanza el número de variables deseado. La eliminación se hace en un bucle de validación cruzada para encontrar el número óptimo de características (Scikit learn, 2019).

* **Entregables y su descripción:**

**HACER GITHUB, o poner donde lo vayamos a entregar**

Los entregables se encuentran en el GITHUB, donde se incluyen tres notebooks en jupyter con el trabajo realizado. El primer notebook incluye el análisis descriptivo, la selección de variables y los modelos sin balanceo de clases y sus métricas. El segundo incluye el balanceo de clases con aumento de minoría, su modelado y métricas. El tercero incluye el balanceo de clases con disminución de mayoría, su modelado y sus métricas. También se incluye el archivo “HR\_attrition\_orig\_proc” en formato csv donde está la base de datos que se utilizó en el estudio. Para la correcta ejecución del código, se necesita instalar las bibliotecas mencionadas previamente, además de almacenar el archivo de la base de datos en la misma ruta que el notebook a ejecutar.

Bibliotecas a instalar en jupyter:

* Matplotlib
* Seaborn
* Scikit-learn
* Pandas
* Numpy
* Statsmodels
* Imblearn
* Collections

También se usan otras librerías preinstaladas.

* **Conclusiones y trabajo futuro:**

En ciencia de datos, es necesario trabajar con cierta metodología. Primero se deben explorar los datos para observar algún hallazgo de interés antes de correr modelos. En caso de tener un número importante de variables, es recomendable hacer feature selection para evitar posibles problemas de multicolinealidad y complejidad dimensional. Es recomendable empezar con el modelo más sencillo e ir explorando hacia el más complejo, y siempre se deben evaluar varios modelos para comparar cuál resuelve mejor el problema.

Cuando se aborda un problema en ciencia de datos, es necesario explorar distintas formas de hacerlo, por ejemplo, en el caso de este trabajo, se balancearon las clases luego de la realización de los modelos iniciales y con tal procedimiento se mejoraron los resultados. También es importante observar cómo se han resuelto problemas similares para abordarlos de la misma manera y comprobar si también funcionan para el problema actual.

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.

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, o incluso la deserción en los cursos online.

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 5. Plan de ejecución original

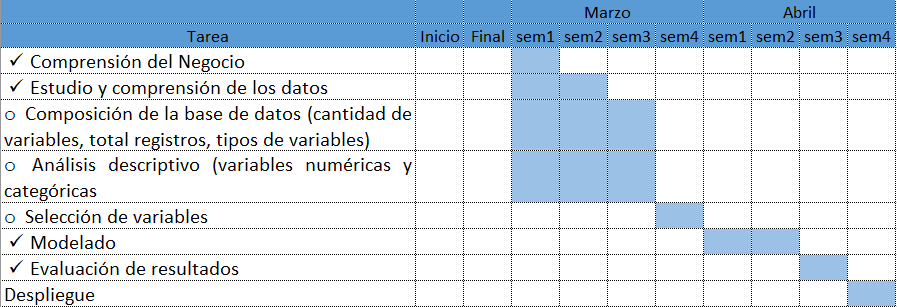


Tabla 6. Plan de ejecución real

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | 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 |  |  |  |  |  |  |  |  |  |  |  |  |

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.

* **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>