PROYECTO GLOBAL

MODELO DE FUGA DE COLABORADORES DE PERFIL DIGITAL DE LA EMPRESA MIBANCO PERÚ

ALUMNO 1: José Antonio Álvarez Guerrero

ALUMNO 2: Ferney Santiago Cardozo Pineda

ALUMNO 3: Robinson Josué Mayta Calderón

ALUMNO 4: Luis David Rodríguez Díaz

ALUMNO 5: Juan Gerardo Rodríguez Monrroy

PROGRAMA:

MÁSTER EN DATA SCIENCE Y BIG DATA

NOMBRE DEL PROYECTO: Modelo de fuga de colaboradores de perfil digital de la empresa Mibanco Perú.

AGRADECIMIENTOS

Ferney Santiago Cardozo Pineda

Quiero agradecer, en primer lugar, a mis padres, por su apoyo incondicional en cada paso que he dado. Sin importar lo arriesgado o novedoso del camino, siempre me brindaron la confianza y el ánimo para seguir adelante, y gracias a ello, encontré lo que realmente me apasiona.

Agradezco también a mi equipo de trabajo, un grupo de personas profesionales, comprometidas e inspiradoras, con quienes he compartido momentos de aprendizaje y crecimiento. A la escuela IEBS, por su excelencia y por abrirme las puertas a un futuro lleno de oportunidades. Y, sobre todo, a Dios, por darme la fortaleza, la humildad y la capacidad de seguir aprendiendo día a día. Hoy me siento más preparado y decidido que nunca a alcanzar mis metas.

Robinson Josué Mayta Calderón

A mis queridos padres, quienes siempre han sido mi fuente de inspiración y fortaleza. En especial, a mi madre, quien en vida abrazó mis sueños con sacrificio y amor. Mamá, tu recuerdo vive en cada paso que doy y en cada logro que alcanzo. Gracias por presentarme a Cristo con tu vida y ser pieza fundamental en todo lo que soy.

A mis hermanos, por su cariño y por ser mi pilar en los momentos difíciles. Juntos hemos superado muchos desafíos y cada uno de ustedes ha contribuido a mi crecimiento personal y académico.

A mi novia, por su amor incondicional y su apoyo constante. Eres mi compañera, mi confidente y mi mayor motivación. Gracias por estar a mi lado en cada momento de este viaje.

A mis compañeros de estudio, con quienes he compartido la experiencia de realizar este trabajo, invirtiendo horas de constante aprendizaje. Gracias por su colaboración, su amistad y por hacer de este máster algo inolvidable. ¡Que lo profesional y la amistad nos lleve a conocernos en un futuro no muy lejano!

Juan Gerardo Rodriguez Monrroy

Quiero agradecer a mis padres, quienes desde el inicio que les conté sobre estudiar una maestría me apoyaron, como lo han hecho desde siempre y con quienes siempre seguiré agradecidos por todo lo que me han ayudado. A mi tía, quien al igual que mis padres, siempre ha estado ahí para ayudarme y para alentarme en seguir adelante. A mi hermana, que al igual que yo tiene muchas metas en las que está trabajando. A mi novia que siempre me apoyó, en especial en momentos donde las cosas a veces se ponían complejas. Y finalmente a mis compañeros, quienes son unas personas muy profesionales, capaces y con quienes estoy seguro de que el trabajo realizado, ha sido de muy alta calidad.

Muchas gracias a todos.

José Antonio Álvarez Guerrero

A mi madre, pues desde siempre nos hemos inspirado el uno al otro, ella apoyándome en mi máster, y yo apoyándola en sus estudios actuales de universidad.

A mi abuela, que al empezar el máster siempre fue feliz de que saliera adelante, y ahora que ya no está con nosotros, estoy seguro de que donde quiera que se encuentre, lo sigue estando por mí.

Finalmente, a mis compañeros de equipo, pues he aprendido de ellos no solo de temas del máster, sino también de las culturas de Latinoamérica, por lo que he desarrollado una gran amistad con ellos y una gran relación en lo profesional, con cada uno de nosotros aportando nuestros valiosos conocimientos.

Contenido

RESUMEN

El problema central de este proyecto radica en el análisis de diversos indicadores que reflejan el comportamiento de los colaboradores con perfil digital de la empresa MiBanco, para ello, se analizarán diversos indicadores basados en datos históricos de los empleados, recopilados entre 2019 y 2024

La base de datos incluye tanto características cualitativas (estado civil, edad, grupo generacional, rango salarial, puesto, grado salarial, sexo) como cuantitativas (sueldo, distancia al trabajo, tiempo en el puesto, tiempo en la empresa, entre otros), los cuales tiene como eje principal al colaborador.

El proceso metodológico que realizamos comprende la depuración de datos para eliminar inconsistencias y valores atípicos, un análisis exploratorio de datos (EDA) utilizando el entorno de distribución Anaconda y cuadernos de Jupyter en Python para identificar patrones y relaciones significativas entre las variables, la construcción de un modelo predictivo aplicando técnicas de machine learning supervisado para predecir la probabilidad de fuga de los colaboradores, y la validación y evaluación del modelo para asegurar su precisión y efectividad utilizando métricas de rendimiento adecuadas.

El objetivo principal es identificar a los empleados con mayor riesgo de fuga, permitiendo a MiBanco implementar estrategias de retención específicas y efectivas, lo que contribuirá a reducir la rotación de personal y a mejorar la satisfacción y el compromiso de los colaboradores con este perfil.

INTRODUCCIÓN

El presente proyecto se centra en el desarrollo de un modelo predictivo para identificar la probabilidad de fuga de colaboradores con perfil digital en la empresa MiBanco. Este análisis se basa en la evaluación de diversos indicadores derivados de datos históricos de los empleados, recopilados entre los años 2019 y 2024. La base de datos incluye tanto características cualitativas (estado civil, edad, grupo generacional, rango salarial, puesto, grado salarial, sexo) como cuantitativas (sueldo, distancia al trabajo, tiempo en el puesto, tiempo en la empresa, entre otros), teniendo al colaborador como eje principal.

La alta rotación de personal en perfiles digitales representa un desafío significativo para las empresas modernas, especialmente en el sector financiero, ya que esto les permite mantenerse competitivas, seguras y eficientes, al tiempo que ofrecen un servicio de alta calidad. MiBanco ha identificado la necesidad de abordar este problema mediante el uso de técnicas avanzadas de análisis de datos y machine learning.

La implementación de un modelo predictivo permitirá a la empresa anticipar la fuga de talento y desarrollar estrategias de retención más específicas y efectivas. Esto contribuirá no solo a reducir la rotación de personal, sino también a mejorar la satisfacción y el compromiso de los colaboradores con perfil digital. Esto contribuirá no solo a reducir la rotación de personal, sino también a mejorar la satisfacción y el compromiso de los colaboradores de perfil digital con el banco.

Para ello el proceso metodológico que realizamos comprende la depuración de datos para eliminar inconsistencias y valores atípicos que existe, permitiendo realizar un análisis exploratorio de datos (EDA) que tiene el banco, identificando patrones y relaciones significativas, a fin de desarrollar un modelo predictivo utilizando técnicas de machine

learning supervisado, y sometiendo el modelo a una validación y evaluación asegurando su precisión y efectividad.

Analizando las bases de datos de MiBanco y debido a que es un problema de regresión logística (modelo de aprendizaje supervisado), con variables tanto categóricas como continuas se determinó aplicar el algoritmo de regresión logística con xGBoost, por ser el mejor para el caso ya que generalmente maneja bien los conjuntos de datos con este tipo de particularidades, donde dependiendo de los parámetros configurados, este algoritmo minimiza la posibilidad de sobre ajustarse al incorporar múltiples árboles de decisión en paralelo y tomar el valor más frecuente de los resultados, es decir evita o reduce la posibilidad de ajustarse al conjunto de datos de entrenamiento y no de forma generalizada, perdiendo precisión al recibir algún dato diferente al que fue entrenado el modelo.

Para el desarrollo de la solución planteada de MiBanco, con el uso de técnicas de Machine Learning para el manejo de los datos, se inició con el proceso de análisis empleando y ejecutando código en el entorno de Python, importando las librerías necesarias para el manejo de dataframes, estadísticas y gráficas para visualización de resultados; permitiendo ver los valores de los conjuntos de datos de train (datos de entrenamiento), test (datos de prueba), cesados (datos de los colaboradores cesados), etc., efectuando un tratamiento a los formatos de los datos, convirtiendo los mismos a los adecuados, y se realizaron uniones de las bases de datos para ejecutar los análisis y predicciones de forma precisa.

Dentro del tratamiento de los datos, se efectúa un análisis estadístico de las bases para considerar e identificar información clave o trascendente; realizando un muestreo descriptivo presentando gráficos de las variables suministradas, análisis de los outliers (datos

anómalos), entre otros. Para ello, se utilizó la librería “shap”, la cual nos permite clasificar las variables más importantes a la hora de intentar predecir el valor de la variable objetivo.

Después de ello, con el tratamiento de los datos y los análisis obtenidos según cada paso aplicado se decidió realizar un modelo de predicción de Regresión Logística, con este mismo modelo se utiliza el método de eliminación de rasgos recursivos para identificar también cuáles son las variables más relevantes.

Con esto y como punto final se obtiene un archivo en Python con la siguiente estructura que permite el análisis y la predicción de los datos:

- Carga de librerías para el desarrollo del modelo
- Carga de bases de datos suministradas
- Análisis de bases de datos
- Depuración de bases de datos.
- Análisis de indicadores sobre las bases de datos.
- Análisis de correlaciones.
- Análisis de los factores más importantes para utilizar en el modelo de predicción.
- Análisis y ejecución del modelo de predicción con diferentes factores.
- Optimización del mejor modelo identificado
- Ejecución y desarrollo del gráfico sobre el modelo desarrollado
- Extracción o generación de los datos en formato Excel.

Se espera que el modelo predictivo desarrollado permita a MiBanco identificar con precisión a los empleados con mayor riesgo de fuga. Esto facilitará la implementación de estrategias de retención más efectivas, mejorando la satisfacción y el compromiso de los colaboradores y reduciendo la rotación de personal.

ESTADO DEL ARTE

Desde hace ya mucho tiempo se ha considerado que las empresas están para generar dinero, es por ello que siempre buscan la manera de poder maximizar las ganancias de cualquier forma.

Una de ellas es tratar de reducir la rotación de personal, la cual genera muchos problemas para las empresas, desde altos costos, inversión de tiempo para contratar nuevo personal, fuga de talento cualificado y demás. Es por ello que, con ayuda de modelos de predicción, big data y demás se plantea el poder detectar cuando una persona está por renunciar a su puesto de trabajo. En específico lo que nos interesa son los perfiles tecnológicos.

1. ***Predicción de renuncia voluntaria de colaboradores con perfil tecnológico de una entidad financiera utilizando regresión logística binaria*:** Es importante que las empresas traten de retener a los perfiles tecnológicos, ya que estos en los últimos años han tomado demasiada relevancia e importancia para que una empresa pueda crecer y sea competitiva contra sus adversarios. En la fuente 1 trataron de reducir el indicador de rotación trimestral en 1.5% con la ayuda de un modelo de regresión logística.

2. ***Implementación de machine learning en la rotación de personal*:** La rotación de personal puede llegar a afectar la productividad, innovación y rentabilidad de una empresa. Siguiendo una solución similar a esta problemática, se ayudaron del análisis de datos y machine learning para poder predecir cuando una persona está a punto de renunciar. Adicional a esto, pudieron explorar ciertas estrategias para aumentar la retención y minimizar la rotación.

3. ***Predicción de fuga de trabajadores utilizando minería de datos:*** Para poder lograr modelos que tengan un nivel alto de predicciones se pueden utilizar metodologías como la LTDM, que permite volver a etapas anteriores para ser modificadas y analizadas en el modelo que se está construyendo.

4. ***Sistema web para la gestión y retención del talento de una empresa outsourcing de TI basado en el aprendizaje automático***: Para ciertos sectores empresariales como el outsourcing, tener una alta rotación de personal puede llegar incluso a generar desconfianza a sus clientes. Una de las claves para poder generar un buen modelo de predicción es el poder detectar qué variables son las que mayormente afectan a la decisión de renunciar por parte de una persona. Ya que una vez que se detectan es cuando se podrán tomar verdaderas acciones por parte del departamento de Recursos Humanos.

5. ***El impacto de la analítica predictiva y prescriptiva en la retención del talento humano en las organizaciones:*** Otra forma en que se ha tratado de dar solución a la rotación de personal es a través de cómo impacta la analítica descriptiva y prescriptiva en ello.

6. ***Rotación de personal. Predicción con modelo de regresión logística multinivel:*** Hablando ya un poco sobre las variables que afectan a que una persona busque renunciar de su empresa podemos mencionar el estado civil, salario y beneficios adicionales por área, estas variables fueron las descubiertas por el trabajo de investigación mencionado anteriormente que utilizó un modelo de regresión logística multinivel para predecir la rotación.

7. ***Aprendizaje automático para la gestión y retención de talento en las organizaciones*:** El estudio de las bajas que hay en una empresa se hace con un enfoque más matemático, al incluir procesos estocásticos y sus bases como lo son las cadenas de Márkov, pues viéndolo sutilmente, el tiempo también es una variable que influye en la decisión de los empleados, recordando que las cadenas de Márkov son procesos cuya probabilidad de que ocurra un evento depende de otro anterior en el tiempo.

8. ***Estudio de caso sobre la problemática de la alta rotación voluntaria del personal en el área de compras en la empresa Equans Perú S.A:*** Otras variables que pueden generar una alta rotación del personal son la inadecuada distribución de labores y el déficit de personal que genera estrés y cansancio.

9. ***Modelo predictivo para la gestión de la rotación del talento humano en una compañía de software:*** Para poder generar un buen modelo de predicción de cuando una persona está a punto de renunciar es factible y aconsejable experimentar con diversos modelos, cuya fuente lo hizo al comparar modelos de regresión logística, ANN y SVM, donde para ellos, el modelo de regresión logística fue el mejor. Similar a otros proyectos, logran generar estrategias para cuando esté a punto de suceder alguna baja de empleados.

10. ***Identificación y análisis de las variables externas e internas que influyen en la estimación de la probabilidad de rotación de empleados:*** Existe un proyecto que en pleno 2024 ha sido publicado, considerando las propiedades de cada una de las variables que influyen en la rotación de empleados de una empresa de Colombia.

11. ***Reducción en la rotación de consultores mediante rediseño de procesos de gestión de personal en SII Group Chile:*** Para finalizar, en una empresa de Chile se detectó que existe una alta rotación de personal, por lo cual, utilizando modelos descriptivos y analíticos, se busca predecirse con la mayor precisión posible y de esa manera implementar planes de mejora para conservar a dichos empleados.

OBJETIVOS

Objetivo general: Generar un modelo de predicción que ayude a la división de R.R.H.H a saber cuándo una persona de áreas digitales esté a punto de renunciar en Mi Banco Perú.

Objetivos Específicos:

1. Analizar y depurar los datos históricos de los empleados de MiBanco, recopilados entre 2019 y 2024, para asegurar la calidad y consistencia de la información utilizada en el modelo predictivo.

2. Realizar un análisis exploratorio de datos (EDA) para identificar patrones y relaciones significativas entre las variables cualitativas y cuantitativas que influyen en la fuga de colaboradores con perfil digital.

3. Desarrollar, validar y evaluar un modelo predictivo utilizando técnicas de machine learning supervisado, que permita predecir con precisión la probabilidad de fuga de los colaboradores, y proporcionar recomendaciones para la implementación de estrategias de retención específicas y efectivas.

4. Medir la rentabilidad de los empleados de MiBanco para evaluar el tipo de perfil de aquellos que han sido cesados y aquellos que el modelo pueda predecir su cese para evaluar técnicas de retención de los empleados más eficientes en términos de rentabilidad.

SOLUCIÓN PLANTEADA

Descripción de variables:

El objetivo del proyecto es poder crear un modelo que pueda predecir en qué momento un empleado que pertenece a puestos tecnológicos cesará de trabajar con la empresa para poder afrontar la problemática de ver una alta rotación en puestos tecnológicos ya que esto puede restar confiabilidad con los clientes o a quienes se les presta servicio. Se ha recolectado una base de datos que nos incluye una lista del personal activo y el personal de cesados en un periodo de 6 años (2019-2024), esto con la intención de obtener los registros de todos los empleados que pertenecen a los puestos tecnológicos y obtener datos valiosos para nuestro modelo de predicción. Ha sido importante que en la extracción de los datos se hayan recolectado la mayor cantidad de variables relevantes e importantes.

Dentro de la lista de activos encontramos las siguientes variables:

S_PERIODO	PERIODO EN EL QUE EL TRABAJADOR SE ENCONTRÓ ACTIVO
S_MAT	CÓDIGO ÚNICO DEL TRABAJADOR DETERMINADO POR EL BANCO
S_GENERO	GÉNERO DEL TRABAJADOR
D_FEC_NACI	FECHA DE NACIMIENTO DEL TRABAJADOR
N_EDAD	EDAD DEL TRABAJADOR
S_GRP_GENERACION	GRUPO GENERACIONAL DEL TRABAJADOR
S_EST_CIVIL	ESTADO CIVIL DEL TRABAJADOR
S_TIP_CTRTO	TIPO DE CONTRATO DEL TRABAJADOR
D_FEC_CTRTO	FECHA DE CONTRATO DEL TRABAJADOR
S_JRND_LAB	JORNADA LABORAL DEL TRABAJADOR
S_GRD_SALARIAL	GRADO SALARIAL DEL TRABAJADOR
S_PUESTO	PUESTO EN EL QUE SE ENCUENTRA AL CIERRE DEL PERIODO EL TRABAJADOR
D_FEC_INI_PST	FECHA EN LA QUE INICIÓ EN EL PUESTO EL TRABAJADOR
N_TPO_PSTO	TIEMPO QUE ESTUVO EL TRABAJADO EN EL PUESTO
S_DIVISION	DIVISIÓN A LA QUE PERTENECE EL TRABAJADOR

S_AGENCIA	AGENCIA A LA QUE PERTENECE EL TRABAJADOR
S_RED_STAFF	FLAG QUE DETERMINA SI ES REDD O STAFF EL TRABAJADOR, DONDE LA RED SON LOS TRABAJADORES QUE TRABAJAN EN AGENCIA Y STAFF SON LOS ADMINISTRATIVOS

Y dentro de la lista de cesados encontramos las siguientes variables:

S_PERIODO	PERIODO EN EL QUE SE DIO EL CESE
S_MAT	MATRÍCULA
D_FEC_CESE	FECHA DE CESE
S_DEC_CESE	DECISIÓN DEL CESE: COLABORADOR O BANCO
S_MTV_CESE	MOTIVO DE CESE DEL SISTEMA
S_SUB_MTV_CESE	SUBMOTIVO DEL SISTEMA
S_MTV_CESE_CAT	MOTIVO DE CESE CATEGORIZADO
S_SUB_MTV_CESE_CAT	SUBMOTIVO CATEGORIZADO
S_OBS_CESE	OBSERVACIÓN DEL CESE REGISTRADA EN EL SISTEMA

Limpieza de datos

Una vez que se cuenta con nuestras bases de datos, se optó por hacer la limpieza con Python en Google Colab ya que brinda oportunidad como equipo de trabajo de acceder al notebook y poder hacerle seguimiento al proceso de limpieza.

Luego de cargar las librerías necesarias se leen las bases de datos primeramente de los empleados activos en una sola base de datos unificada que contiene todos los años, seguido a esto lo primero que hicimos como equipo es revisar lo que tenemos, la relevancia de las variables, si existen valores nulos en las variables, la cantidad de registros totales, los registros repetidos, etcétera.

Seguido a ello, se identificó que la base contiene empleados que pertenecen a distintos departamentos o con empleos que no son acorde a nuestro objetivo, el cual es trabajar con todos aquellos empleados que pertenezcan a puestos tecnológicos.

Comenzamos la limpieza haciendo este primer filtro, quedándonos con todos los registros que pertenezcan a puestos tecnológicos; este filtrado lo llevamos a cabo por palabras clave en los puestos de trabajo

y por división, así eliminando todos aquellos registros de empleados que no pertenezcan al nicho que necesitamos. Cabe recalcar que hemos comenzado con alrededor de 600 mil registros y a lo largo de la limpieza realizada, estos registros disminuyen en cantidad, pero aumentan en relevancia y valor para el desarrollo de nuestro modelo.

El segundo filtro que llevamos a cabo fue el de eliminar a todos los empleados que en la última columna pertenecieran a "la red", pues al ser un banco, los empleados que tenían esta categoría eran aquellos que trabajaban en call centers, recordando que nuestra meta principal es trabajar con empleados que se encuentran en corporativo en áreas tecnológicas, por lo que, al hacer esta limpieza, nos quedamos con poco más de 190 mil registros (una reducción de casi el 70% total de la base unificada).

Luego, a través de la limpieza se ha verificado en cada una de las variables si existían nulos, y de esta manera poder reemplazarlos o eliminarlos, pero siempre se decidió buscar ponerles algún valor ya que necesitábamos la mayor cantidad de datos importantes y relevantes, para la limpieza de obtener los empleados activos que pertenecían a puestos tecnológicos tuvimos varias fases de prueba donde filtramos por palabras clave, luego por división, después por departamentos, y de esa forma en cada fase verificamos si existían valores nulos, hasta quedarnos con los registros que pertenecían a puestos tecnológicos, y en muchos casos se tuvo que observar uno a uno cada puesto debido a su poca frecuencia y así con el mayor cuidado no saltarnos registros importantes y relevantes.

A este punto descartamos todos aquellos trabajos, divisiones, empleados que no pertenecen a nuestro nicho para así concentrarnos en verificar el peso e importancia de los datos útiles que teníamos.

Es aquí donde identificamos problemas como el notar que los datos de un mismo empleado se pueden repetir si lleva trabajando varios años ya que se repite su matrícula cada periodo (cada periodo abarca un mes del año) y si tuvo cambio de puesto también, así que tuvimos varios empleados repetidos, pero con diferentes tiempos, puestos de trabajos, rangos salariales, y nuestro primer objetivo era poder unificar todo y obtener un solo registro por empleado y obtener la mayor cantidad de información importante.

Para esto se decidió comenzar a contar la cantidad de puestos que ha tenido cada empleado, el tiempo que lleva en la empresa, contando su primer y último registro que nos aparece. A medida que realizamos

estas limpiezas nos encontrábamos con otros problemas tales como definir un rango salarial, el estado civil, el tipo de contrato, ya que a veces no había una tendencia, así que se decidió definir un valor a los rangos salariales y asignarlos aleatoriamente dentro del rango que tuviera cada empleado y esto a su vez era determinado por el tipo de puesto que tenía, a definir por tendencia y asignando aleatoriamente el tipo de contrato, el estado civil, entre otras.

Eliminamos las columnas que se comenzaban a volver redundantes o que ya no eran relevantes para la limpieza (las cuales fueron: **D_FEC_NACI, S_AGENCIA** y **S_RED_STAFF**) y a agregar columnas que resuman o definan la información con mayor claridad, siempre verificando que no quedaran valores nulos o pesos en las variables fuera de lo común. Al finalizar de hacer esta limpieza se pudieron unir los datos repetidos de cada empleado por su matrícula, y se obtuvo las variables por registro de:

- Periodo
- Matrícula
- Género
- Edad
- Generación
- Estado Civil
- Tipo de Contrato
- Fecha de primer registro
- Jornada Laboral
- Puesto
- Última Fecha Registrada
- Número de meses en el último puesto
- División
- Cantidad de Puestos
- Rango salarial

Con dichas columnas, y aplicando los filtros anteriores, obtuvimos un total de 2249 registros útiles y sin nulos que son aprovechables para nuestro modelo, los cuales, cada registro es un empleado único. Luego de ello, hacemos un barrido final de nulos, columnas relevantes, valores sin sentido para ya poder dejarla lista, por último, antes de hacer la unión con la base de cesados, se hace un renombrado de las columnas para mayor entendimiento.

Después, cargamos la base de cesados y el primer filtro que hacemos para poder determinar los empleados que se fueron y pertenecen a puestos tecnológicos es verificar cuales de nuestra base de datos de

activos aparecen en la base de cesados, y de esta manera nos quedamos solo con los cesados de puestos tecnológicos. Eliminamos columnas irrelevantes, o redundantes y hacemos la unión con la base de datos de activos.

Sabíamos que con el tiempo se tenía que determinar una variable target que sería si es cesado o no para poder realizar nuestro modelo, así que primero hicimos arreglos de las variables ya que sabíamos que aparecerían nulos en aquellos empleados que aún seguían activos, pero no tenían que haber nulos en los empleados ya cesados, así que nos tomamos la tarea de verificar columna por columna, los nulos, el motivo del cese y categorizar para hacer que nuestro modelo pueda identificar mejor.

De esta forma, todos aquellos cesados se le adicionó información que nos indicaba por qué cesaron, si renunciaron o fueron despedidos y la razón real del por qué sucedió además de la fecha en que se hizo el cese ya que con esta fecha podríamos determinar cuánto tiempo duro en la empresa, cuántos cambios de puesto tuvo, cuantos cambios de salario, que tipos de contratos tuvo, si hubo cambios en su información personal, ya que cada variable le agregaba peso a nuestro modelo y así pudimos agregar una variable de cesado donde aquel que no era cesado tenía un '0' y quien si era cesado tenía un '1' así el target era claro. De esa forma, la variable target es binaria, por la cual, podíamos comenzar con nuestro modelado.

Así mismo, decidimos agregar una columna llamada "rentabilidad", pues además de generar un modelo capaz de predecir si un empleado del área tecnológica va a dejar la empresa, queremos ver si existen empleados cuya rentabilidad es alta y aun así dejan el banco, o bien, si existe algún patrón de ciertas áreas en específico que sean más susceptibles a abandonar sus puestos. Dicha rentabilidad la calculamos como un porcentaje definido por el rango entre el salario inicial y final del empleado, multiplicado por el total de puestos que ha tenido y dividido entre el costo total de dicho empleado. Todo esto multiplicado por 100.

De esta manera quedamos con una base de datos final donde ya no había valores nulos y cada una de las variables aporta un peso e importancia para comprender el cese de los empleados en puestos tecnológicos, y así, definir la base final lista para el siguiente proceso que era crear el modelo predictivo

Por último, se decide subir las bases de datos por seguridad a AWS y darles acceso a los colaboradores sólo a través de un link único, así de esta manera se podían eliminar las copias locales que se podrían haber hecho del equipo de trabajo y cuidar la información que tenemos.

El enlace al código de limpieza se encuentra en los anexos.

EVALUACIÓN

Para la realización del modelo se comenzó por la importación de las librerías que serán necesarias para el desarrollo del análisis y el modelo, se realizó la conexión con drive para poder cargar la base de datos final para nuestro modelo.

```python
import pandas as pd

from google.colab import drive
drive.mount('/content/drive')
```

Figura 1: Librerías importadas en google colab

Cargamos la base de datos final con ayuda de pandas, y comenzaremos con el análisis exploratorio, el propósito de este modelo será saber si una persona será cesada o no de la compañía de acuerdo con los comportamientos previstos por esta.

Comenzando con el análisis exploratorio tanto univariado como multivariado revisando las distribuciones de nuestras variables categóricas usando gráficos de barras e histogramas para las variables numéricas.

```python
import matplotlib.pyplot as plt
import seaborn as sns

for col in data.columns:
  if data[col].dtype == 'object':
    # Gráfico de barras para variables categóricas
    plt.figure()
    sns.countplot(x=col, data=data)
    plt.title(f'Distribución de {col}')
    plt.xticks(rotation=45, ha='right')
    plt.show()
  else:
    # Histograma para variables numéricas
    plt.figure()
    sns.histplot(x=col, data=data, kde=True)
    plt.title(f'Distribución de {col}')
    plt.show()
```

Figura 2: Código para la construcción de gráficas

Figura 3: Gráficas obtenidas del análisis exploratorio

Para observar si existe una correlación entre las variables que utilizamos graficamos la matriz de correlación, para lo que primero codificamos las variables categóricas con ayuda del label *encoder*.

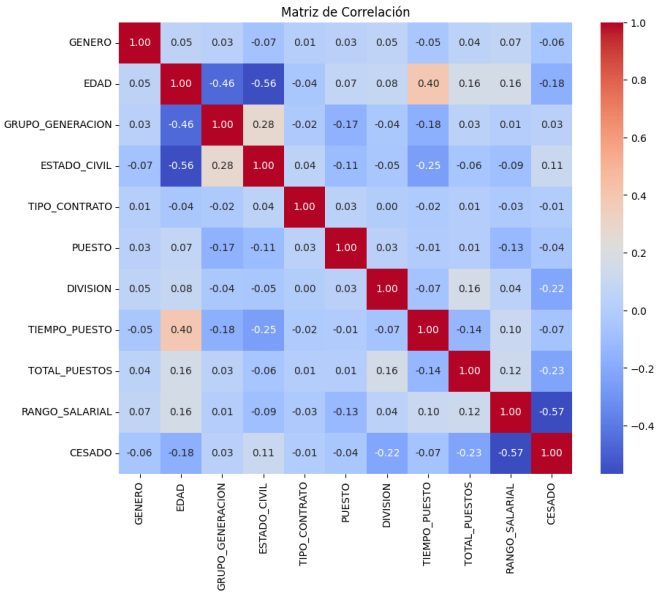

Figura 4: Matriz de correlación de las variables estudiadas

Utilizamos como parámetro umbral +/- .3 y quitamos las variables que están fuera de este parámetro, obteniendo de nuevo la matriz de correlación observamos que solamente nos quedamos con las variables con menor correlación entre ellas , una de ellas nuestra variable que utilizaremos como target, la variable “CESADO”.

Dentro de las variables quitamos la edad ya que se encuentra correlacionada con el tiempo en el puesto y el grupo generacional, así que optamos sólo por mantener el grupo generacional.

Para realizar la selección de variables se utilizó la técnica de *shap values* la cual nos ayudó a indicarnos cuales son las variables que aportan una mayor información al modelo con ayuda de clasificación como es el XGBoostClassifier.

```
import xgboost as xgb
from sklearn.model_selection import train_test_split
import shap

# Dividir los datos en conjuntos de entrenamiento y prueba
X = data_encoded[low_correlation_vars].drop("CESADO", axis=1)
y = data_encoded['CESADO']
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size=0.2,
                                                    random_state=42)

# Entrenar un modelo XGBoost
model = xgb.XGBClassifier()
model.fit(X_train, y_train)

# Explicar el modelo utilizando SHAP
explainer = shap.Explainer(model)
shap_values = explainer(X_test)

# Graficar la importancia de las variables
shap.summary_plot(shap_values, X_test, plot_type="bar")
```

Figura 5: Código con la selección de conjunto de datos, entrenamiento del modelo y técnica shap

A partir de esta técnica logramos observar que las variables más importantes en orden de importancia para la predicción del modelo son:

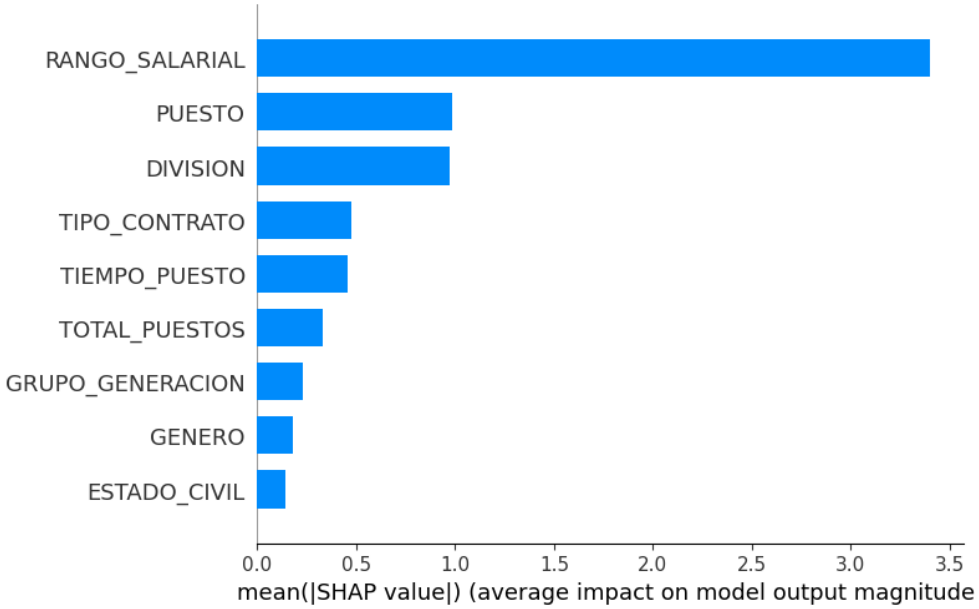

Figura 6: Variables más importantes

Después de este pequeño análisis nos dimos cuenta de que las variables que más aportan información al modelo son

* RANGO SALARIAL

* DIVISIÓN

* TIPO_CONTRATO

* TIEMPO_PUESTO

* GRUPO_GENERACION

* ESTADO_CIVIL

* GÉNERO

* TOTAL_PUESTOS

* PUESTO

RESULTADOS

Comenzamos con el modelado con una distribución 80% y 20% utilizando una regresión logística y un *xgboost* como modelos, utilizando el AUC como medida de precisión del modelo.

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import roc_auc_score

# Seleccionar las variables más importantes
X = data_encoded[["RANGO_SALARIAL", "DIVISION", "PUESTO", "TIEMPO_PUESTO"]]
y = data_encoded['CESADO']

# Dividir los datos en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
                                                    random_state=42)

# Crear y entrenar el modelo de regresión logística
model = LogisticRegression()
model.fit(X_train, y_train)

# Predecir las probabilidades en el conjunto de prueba
y_pred_proba = model.predict_proba(X_test)[:, 1]

# Calcular el AUC
auc = roc_auc_score(y_test, y_pred_proba)
print("AUC:", auc)
```

Figura 7: Código con la creación del modelo de predicción

Se tomaron solo las cuatro variables más influyentes dentro de nuestro análisis, las cuales fueron: "*RANGO_SALARIAL*", "*DIVISION*", "*PUESTO*", "*TIEMPO_PUESTO*".

Obteniendo como resultados un AUC de .875 en test lo cual es una buena métrica, apoyándonos del gráfico siguiente que nos ayuda a saber que nuestro modelo no se encuentra sobre ajustado, por lo que podemos concluir que es un buen modelo de aproximación.

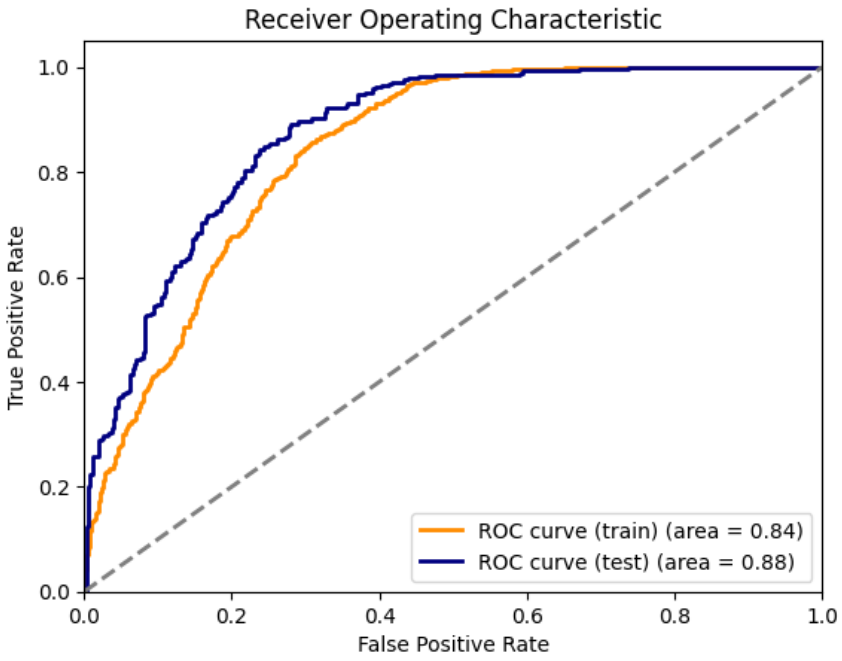

Figura 8: Curva ROC del modelo de predicción

Para la base test, se agregaron columnas de predicción y del valor real, para ver cuáles son las aproximaciones del modelo en cuestión y el valor real de la variable, tomando en cuenta lo siguiente:

Sea x: Evento que describe si un empleado es cesado o no.

P(x): Probabilidad de que ocurra el cese del empleado en cuestión.

Recordando que nuestra variable target es binaria, se intuye lo siguiente:

- Si $P(x) \leq 0.5$, se evalúa la predicción como 0 (el empleado no ha cesado).
- Si $P(x) > 0.5$, se evalúa la predicción como 1 (el empleado ha sido cesado).

	RANGO_SALARIAL	DIVISION	PUESTO	TIEMPO_PUESTO	Prediccion	Target
2210	487	4	689	0	0.539874	1
482	1132	0	262	12	0.270462	0
1804	53	2	283	21	0.945440	1
247	1272	7	199	6	0.050194	0
1656	338	1	423	73	0.830622	0

Figura 9: Resultado del test

Utilizando estos resultados, procedemos a operar con la variable de rentabilidad. Una propuesta que se le puede dar a "MiBanco" es que aquellos empleados que el modelo predice un valor cercano a uno (bajo los criterios mencionados arriba), es tratar de platicar con dichos empleados cuya rentabilidad sea mayor a cierto porcentaje definido entre directivos y recursos humanos.

Aplicando dicho análisis a nuestros resultados, se obtuvo lo siguiente:

```
X_test[X_test["Prediccion"] > 0.5]["RENTABILIDAD"].mean()
```

```
4.842155913038587
```

Figura 10: Resultado de la rentabilidad

El promedio de la rentabilidad de esos empleados con predicción alta a abandonar la empresa es de aproximadamente 5%, lo que indica que todos estos empleados tienen baja rentabilidad para la empresa. Se procede a ver cuántos empleados tienen dichas características.

La base test contiene 441 registros (empleados), de los cuales, 110 tienen una predicción de cese arriba del 50%, lo que corresponde aproximadamente a la cuarta parte de nuestra base final.

Se hace una comprobación de cuántos empleados en realidad dejarían la empresa. Para ello, se agregó una variable "target", la cual contiene el valor correcto de la variable analizada en el modelo. Esto para corroborar la efectividad de nuestro modelo.

```
pruebas[pruebas['Target']>0] #De 110 empleados con predicción cercana a 1, el 80% de ell
```

	RANGO_SALARIAL	DIVISION	PUESTO	TIEMPO_PUESTO	Prediccion	Target	RENTABILIDAD
1161	639	0	263	45	0.710435	1	13.956322
1085	589	0	267	68	0.737546	1	13.328919
2226	307	0	264	56	0.900106	1	12.863128
1714	497	8	63	5	0.537262	1	12.812644
1874	605	4	412	4	0.518199	1	10.038984
...	...	...	...	...	...	...	...
1880	219	2	569	5	0.860456	1	0.421890
2138	184	6	160	3	0.845079	1	0.378695
1972	194	2	295	2	0.910421	1	0.222649
1216	32	8	562	4	0.786137	1	0.116755
1754	425	2	158	11	0.827200	1	0.085174

88 rows × 7 columns

Figura 11: Tabla con los registros de empleados con predicción de cese mayor al 50%

Se puede observar que 88 de 100 empleados (80%) se predicen correctamente, acorde a lo obtenido en la curva ROC. De esos empleados, filtramos solo aquellos con una rentabilidad mayor al 100%, obteniendo como resultado final 14 empleados con esas características.

	RANGO_SALARIAL	DIVISION	PUESTO	TIEMPO_PUESTO	Prediccion	Target	RENTABILIDAD
1486	602	3	360	32	0.574618	0	22.295150
1071	556	6	144	80	0.507850	0	18.764208
1086	617	4	229	88	0.517824	0	17.246901
1161	639	0	263	45	0.710435	1	13.956322
1085	589	0	267	68	0.737546	1	13.328919
1333	416	6	626	0	0.526985	0	13.142083
2226	307	0	264	56	0.900106	1	12.863128
1714	497	8	63	5	0.537262	1	12.812644
1321	598	2	640	4	0.550144	0	12.668831
1258	363	6	622	5	0.577705	0	12.625539
916	431	8	55	22	0.594290	0	12.317296
1170	668	2	159	6	0.646446	0	11.025953
932	607	1	591	8	0.607836	0	10.094062
1874	605	4	412	4	0.518199	1	10.038984

Figura 12: Tabla con 14 empleados con una rentabilidad mayor al 10%

Logramos observar que son casos puntuales, pues además de que son pocos trabajadores, no existe una característica amplia de qué tipo de empleado es el que deja la empresa, pues éstos provienen de varias divisiones, distintos puestos, e incluso algunos con bastante tiempo en la empresa pueden llegar a abandonar al banco.

Para el caso concreto del empleado con valor "cero" en "*TIEMPO_PUESTO*", es un caso atípico de un empleado con un buen rango salarial final que, al momento de corte de nuestras bases de datos (JUNIO 2024), recién fue ascendido de puesto, por lo que al momento de aplicar nuestro análisis, recién había tomado ese puesto en el último periodo. Observemos que esto se comprueba en la variable "*TARGET*", con un valor real de 0 y una predicción apenas por encima del 50%.

El enlace al código que contiene la aplicación del modelo e interpretación de resultados se encuentra en la parte de anexos.

CONCLUSIONES Y TRABAJOS FUTUROS

El modelo de regresión logística nos ha permitido llevar a cabo un análisis breve pero preciso sobre si los empleados de la empresa "MiBanco Perú" de áreas tecnológicas van a abandonar la empresa o no. El resultado obtenido fue un área bajo la curva de 0.87, lo cual nos indica una buena aproximación sin sobreajuste, ya que el conjunto de entrenamiento tuvo un menor valor. Al aplicar las librerías "shap" y "XGboost", pudimos observar que solo ciertas variables tenían mayor correlación con la variable binaria objetivo.

Como resultado, se obtuvieron dos códigos de Python creados en Google colab que contienen la limpieza y depuración de los datos, un análisis exploratorio que nos ayudó a entender aún mejor nuestros datos y un modelo de regresión logística que nos ayudó a predecir con una precisión del 87% cuando un empleado relacionado a dichas áreas está a punto de cesar. Esto también trajo como resultado evaluar cada caso y en base a la rentabilidad de cada uno, se puede determinar el esfuerzo y recursos que "MiBanco Perú" podría usar para retener a esos perfiles.

Dados los objetivos planteados, una vez que se ejecutó el modelo y se obtuvieron las predicciones, se analizó la variable de rentabilidad, esto con el fin de observar si hay algún patrón entre dicha variable y la predicción de su posible cese. Con ello, llegamos a la conclusión de que los empleados menos rentables (aquellos con una rentabilidad negativa) son los más propensos a abandonar su puesto de trabajo, mientras que el valor más alto de rentabilidad de los posibles trabajadores propuestos a renunciar es apenas mayor al 20%.

Una de las limitantes que encontramos a la hora de ejecutar nuestro Global Proyect fueron las variables con las que contamos, ya que eran las disponibles por parte de "MiBanco Perú", ya que tener más variables como la cantidad de proyectos en las que el colaborador ha trabajado, o bien, cuantas promociones de aumento de salario ha tenido en todo su periodo laboral, habrían mejorado el cálculo de la rentabilidad del empleado, ya que solo se tenía la cantidad de veces que ha cambiado de puesto, más no las veces adicionales en que su salario se vio aumentado.

Otra de las limitantes fue la cantidad reducida de investigaciones parecidas a nuestro proyecto, ya que si bien se encontraron algunas que incluso seguían la misma línea de crear un modelo predictivo para

detectar cuando un empleado estaba a punto de renunciar, otras lo hacían con otros métodos como sistemas web o minería de datos.

Como línea de investigación futura y con el avance acelerado del desarrollo de inteligencias artificiales, se podría llegar a tener un sistema a disposición de los equipos de R.R.H.H que determinen de forma aún más precisa el momento en que una persona esté comenzando a tomar la decisión de salir de su empresa.

Como punto final, al tener una precisión muy buena, el modelo podría llevarse a producción a los sistemas de RR. HH. para que el departamento comience con medidas de prevención de ceses. Como casos puntuales, de los 14 empleados con buena rentabilidad que el modelo predijo un posible cese, se podrían implementar planes de mejora a futuro, como una posible renegociación de sueldos, algún bono adicional, salario emocional (recompensas con la familia del empleado), etcétera, para que de esa forma la tasa de empleados cesados por renuncias pueda bajar en el futuro. Así mismo, se podría nutrir las bases de empleados activos y pasivos con más variables que ayuden mejor al modelo, como lo pueden ser la cantidad de proyectos que el empleado ha tenido, o bien, si se le han dado bonificaciones adicionales.

REFERENCIAS

Montoya, R., & Rubén, R. (2023). *Predicción de renuncia voluntaria de colaboradores con perfil tecnológico de una entidad financiera utilizando regresión logística binaria*. Universidad Nacional Agraria La Molina.

Castro Bravo, A. C., Otalora Cubides, L. E., Sánchez Olave, L. K., & Toquica Toquica, P. L. (2024, April 4). *Implementación de machine learning en la rotación de personal*. Repository.universidadean.edu.co. https://repository.universidadean.edu.co/handle/10882/13431

Valdés, P., & Profesor, D. (2021). DSpace Biblioteca Universidad de Talca (v1.5.2): Predicción de fuga de trabajadores utilizando minería de datos. *Utalca.cl*. http://dspace.utalca.cl/handle/1950/13345

Victoria, C., & Agusto, C. (2022). Sistema web para la gestión y retención del talento de una empresa outsourcing de TI basado en el aprendizaje automático. *Upc.edu.pe*. http://hdl.handle.net/10757/660875

Lorusso, A. (2022). El impacto de la analítica predictiva y prescriptiva en la retención del talento humano en las organizaciones. *Ub.edu.ar*. http://repositorio.ub.edu.ar/handle/123456789/9860

De Estadística, E. (n.d.). *UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS FACULTAD DE CIENCIAS MATEMÁTICAS*. Retrieved September 7, 2024, from https://core.ac.uk/download/pdf/323348949.pdf

Frieri, L. M. (2017). Aprendizaje automático para la gestión y retención de talento en las organizaciones. *Repositorio.utdt.edu*. https://repositorio.utdt.edu/handle/20.500.13098/11151

Arturo, D., & Miguel, G. (2021). Estudio de caso sobre la problemática de la alta rotación voluntaria del personal en el área de compras en la empresa Equans Perú S.A., 2021-2023. *Upc.edu.pe*. http://hdl.handle.net/10757/674683

Leonardo. (2024). Modelo predictivo para la gestión de la rotación del talento humano en una compañía de software. *Itm.edu.co*. http://hdl.handle.net/20.500.12622/6570

Hernández, L., & Angel, L. (2024, July 19). *Identificación y análisis de las variables externas e internas que influyen en la estimación de la probabilidad de rotación de empleados*. Repositorio Institucional Séneca; Universidad de los Andes.

https://repositorio.uniandes.edu.co/entities/publication/dca8adf5-343e-4de5-b005-1144b911c4b5

Antonio, M. (2017). Reducción en la rotación de consultores mediante rediseño de procesos de gestión de personal en SII Group Chile. *Uchile.cl*. https://repositorio.uchile.cl/handle/2250/147213

ENLACES

https://repositorio.lamolina.edu.pe/handle/20.500.12996/6228

https://repository.universidadean.edu.co/handle/10882/13431

http://dspace.utalca.cl/handle/1950/13345

https://repositorioacademico.upc.edu.pe/handle/10757/660875

http://repositorio.ub.edu.ar/handle/123456789/9860

https://core.ac.uk/download/pdf/323348949.pdf

https://repositorio.utdt.edu/handle/20.500.13098/11151

https://repositorioacademico.upc.edu.pe/handle/10757/674683

https://repositorio.itm.edu.co/handle/20.500.12622/6570

https://repositorio.uniandes.edu.co/entities/publication/dca8adf5-343e-4de5-b005-1144b911c4b5

https://repositorio.uchile.cl/handle/2250/147213

ANEXOS

Enlace código de limpieza:

https://colab.research.google.com/drive/1HekPqOBp4P-IEK6Hj9t_b80Lctqn1Roh?usp=sharing

Enlace código del modelado:

https://colab.research.google.com/drive/1sA9j-ZF9kV_vJltVYfxCofI8RRD83WIm?usp=drive_link

Drive con bases iniciales (sin transformaciones):

https://drive.google.com/drive/folders/1TGVq3wo2mKSb3BOYTdvgpU9cMkTMrtKf?usp=sharing