TRABAJO FINAL

DS1

MODELO CLASIFICACION: PREDICCION ABANDONO BANCO MACRO

ALUMNO: PABLO SOTOMAYOR

INTRODUCCION:

Banco Macro es una banco argentino que tiene clientes en todas las provincias de Argentina y en tres países europeos (Alemania, España y Francia).

En este proyecto en particular vamos analizar y predecir la probabilidad de que un cliente abandone o no la institución en esos 3 países europeos.

Para ello tomaremos en cuenta la información generada por la administración y diseñada por el equipo de ingeniería de datos del banco, en cuanto a clientes que ya haya abandono o no el banco.

En las siguientes paginas vamos a exponer y explicar todas las etapas del proceso: comenzando con los objetivos y pasando por la carga de datos hasta la verificación del modelo.

DISEÑO CONCEPTUAL:

1- Misión y visión:

La fidelización de los clientes es un punto central de las compañías en la actualidad. Las compañías se desarrollan, existen y se sostienen debido a la existencia de los mismos. Esta situación se profundiza aún más cuando hablamos de instituciones financieras tradicionales, las cuales en los últimos años han enfrentado una competencia sin precedentes (debido a los nuevos formatos de instituciones financieras).
Por tal motivo el análisis y la predicción de la pérdida o deserción de clientes es fundamental para la toma de decisiones, de los bancos tradicionales, en relación a las políticas de retención de cartera (clientes).

Como respuesta a ello, este trabajo se centra en la exploración, análisis y tratamiento de un conjunto de datos relacionados con la deserción de clientes y, en consecuencia, en la construcción de un modelo de clasificación que predice si un cliente abandonará el banco o no.

2- Descripción del problema:

La deserción de clientes es un problema que pone en riesgo las ganancias de la compañía (el banco), así como también, su estabilidad y solvencia. Desarrollar un modelo predictivo clasificatorio, con la calidad suficiente, que arroje diferentes insights, es de suma importancia para la toma de decisiones en relación a la retención de clientes.

3- Temática de los datos y alcance del proyecto:

Para este proyecto se toman los datos generados por el equipo de administración del banco, en donde constan las personas que abandonaron el banco y las que no. Los mismo fueron subidos a Kaggle.

https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset

Este dataset cuenta con información relevante acerca de clientes, características de los mismos y perfil bancario de estos (productos, antigüedad, puntaje crediticio, etc.). Esta información nos permitirá crear un modelo predictivo para determinar la deserción de clientes.

4- Objetivo:

El objetivo primordial de este trabajo es construir un modelo de aprendizaje supervisado clasificatorio, que sea capaz de predecir si un cliente va a abandonar el banco o no.

5- Usuarios finales y nivel de aplicación:

   Este proyecto está dirigido a los miembros de los equipos de markenting y fidelización de clientes. A través de la predicción de los clientes que abandonarán el banco o no los mismos pueden tomar acciones para la retención de estos últimos (clientes).

HIPOTESIS:

1- Primeros pasos para llegar a la hipótesis:

   Primero importamos librerías de python, luego leímos el csv, ordenamos el mismo y realizamos las primeras visualizaciones para corroborar que no haya errores en la importación.

2- Exploración de datos:

   Ahondamos más en los datos, visualizamos variables que serían relevantes en nuestro trabajo, divisamos la variable a predecir (si abandono o no el cliente) y sacamos nuestras primeras conclusiones obteniendo datos estadísticos de los datos.

3- Preguntas de investigación:

   ¿Cuáles son las variables que se relacionan con la variable de abandono? ¿Qué relación existe entre esas variables y la variable abandono? ¿incide el país de origen o el sexo con el abandono? ¿incide el puntaje crediticio o el tiempo que el cliente lleva en el banco en el abandono? ¿o quizás el saldo de cuenta, el salario del cliente o la cantidad de productos que tenga el mismo? ¿Que otra cosa podría incidir en el abandono?

4- Hipótesis:

A través de las variables mencionadas anteriormente se puede predecir si un cliente abandonará el banco o no.

1. Existe una diferencia en cuanto al nivel de deserción de acuerdo al país del cliente.
2. Existe una diferencia en cuanto al nivel de deserción de acuerdo al género del cliente.
3. El abandono del banco varía linealmente en relación con el puntaje crediticio: a menor credit_score más deserción.
4. Cuanto mayor es el tiempo de permanencia menor es la deserción.
5. La deserción aumenta cuando disminuye el saldo de la cuenta: a menor saldo más abandono.
6. A menor cantidad de productos mayor deserción. Es decir, la permanencia aumenta con la cantidad de productos.

VISUALIZACIÓN Y EXPLORACIÓN DE LAS HIPOTESIS CON GRAFICOS:

Por cada una de las hipótesis establecimos la relación de la variable involucrada con el abandono de los clientes. Se gráfico en, cada caso, con las herramientas adecuadas y obtuvimos la relación que buscábamos.

En países por ejemplo Alemania tuvo el mayor porcentaje de deserciones.

PREPROCESAMIENTO DE DATOS:

En esta etapa del proceso normalizamos y limpiamos los datos para que el modelo lo ingeste de manera adecuada.

Es decir, limpiamos los datos para que el modelo sea eficiente y funcione correctamente.

Tratamos y corregimos los valores extremos (outliers) para que no incidan negativamente en el entrenamiento del modelo, manejamos los valores nulos o faltantes con diversas técnicas, se normalizan los datos que presenten diferencias, eliminamos las variables que no son relevantes para el modelo y codificamos las variables categóricas para que el modelo pueda procesarlas. Es decir, se convierten las variables categóricas como sexo o país en variables numéricas.

En nuestro caso también corregimos los datos con datos sintéticos ya que el conjunto de datos no lograba ser suficientemente representativo. Esto hizo que el modelo tenga mayor asertividad.

CONSTRUCCION DE MODELOS:

Acá se construyen los modelos con los datos refinados y preprocesados. Se entrena el mismo y se testea.

Nosotros elegimos tres modelos para obtener una mayor certeza y poder comparar los datos, las predicciones y las métricas que arroja cada uno.

Se dividen los datos en dos partes. En la primera parte nosotros dejamos el 70 por ciento de la muestra y esta sirve para entrenar el modelo. En la segunda parte dejamos el 30 por ciento de los datos y sirve para testear el modelo y entender si el mismo fue bien entrenado.

VALIDACION DE MODELOS:

Acá vamos a validar nuestros modelos con las métricas más utilizadas para los modelos de clasificación, incluyendo a la matriz de confusión.

En esta validación guardamos los resultados de la predicción que hace el modelo (de la variable a predecir), tomamos los mismos junto a los datos de testeo de la misma variable y comparamos el grado de asertividad de la predicción utilizando varias métricas estándar para la industria.

Es decir, comparamos si la predicción se correlaciona con los datos reales.

Si las métricas nos devuelven que no hay un grado importante de asertividad se toman las medidas necesarias, a través de ciertas técnicas, para corregir el dataset y entrenar el modelo de manera eficiente.

En nuestro caso el modelo tiene un grado de asertividad alto por que tomamos las medidas adecuadas para corregir desvios.

CONCLUSION:

Este modelo de clasificación es sumamente importante para predecir si un cliente va a abandonar el banco. Esto nos permite adelantarnos y aplicar todas las estrategias de negocio para la retención y fidelización de esos clientes.

Entendiendo quienes tienen mayor probabilidad de deserción se puede accionar sobre esos clientes para ofrecerle distintos productos y beneficios para su fidelización.

En un mercado financiero cada vez más competitivo esto es fundamental.