IEEE IEMDC

IEEE INTERNATIONAL ELECTRIC MACHINES AND DRIVES CONFERENCE

June 1-4, 2003

Madison, Wisconsin

Segmentación de usuarios y marketing dirigido.

Germán Wong del Toro, Héctor Miranda García, Victor Hugo Portilla Ortiz, Misael Chavez Ramos

Diego López Bernal
David Balderas
Jesús Manuel Vázquez Nicolás, Dr. Oscar Fuentes Covarrubias, Emanuel Páez López, Andrea Torres

Introducción

La pérdida de clientes por parte de las empresas es un evento inevitable. Esto puede ser provocado por diversas razones, tales como una competencia, insatisfacción por parte de los productos o servicios, cambios en las preferencias de los clientes o experiencias negativas que han tenido. Debido a estos diferentes factores, se busca una solución con la implementación de métodos computacionales para obtener una solución efectiva del problema.

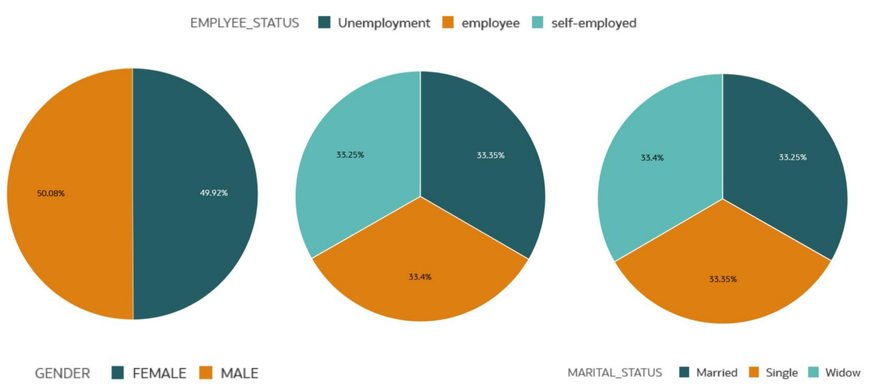
Objetivo

Nuestro objetivo es emplear herramientas computacionales para llevar a cabo un análisis de datos de los clientes y obtener una segmentación de los mismos, con el fin de identificar las posibles causas por las que están abandonando la empresa. Los datos recopilados se demostrarían por medio de una plataforma web con el objetivo de proporcionar al equipo de marketing un medio para realizar marketing dirigido.

Análisis de datos

Previo al desarrollo de los modelos y de la ingeniería de atributos, se hizo un análisis de los datos proporcionados. Este análisis constó de observar diferentes métricas para cada atributo, así como visualizaciones de distribución, correlación, etc. para tratar de identificar comportamientos interesantes o indeseados.

Encontramos varios detalles que dificultaron este paso, ya que nuestro dataset parece ser sintético. Estos detalles incluyen un balance casi perfecto de atributos categóricos y la existencia de multicolinealidad entre atributos.



De igual manera, encontramos diversos errores en el dataset, como llaves primarias duplicadas, y datos perdidos por una codificación defectuosa.

Preprocesamiento

Implementamos una serie de pipelines para facilitar el tratamiento de los datos, tanto a la hora de entrenar como para la predicción.

En general, eliminamos atributos que consideramos irrelevantes para el enfoque de nuestro proyecto, así como atributos multicolineales, ya que estos afectan la gran mayoría de modelos propuestos. De igual manera, se imputaron datos faltantes o erróneos, como los mencionados anteriormente, con ayuda de LLMs. Por último, se usaron diferentes métodos de encoding para atributos categóricos dependiendo de la cantidad de clases.

Para el caso del modelo no supervisado de clustering, decidimos tratar la gran mayoría de outliers usando rangos intercuartiles. Este método nos garantiza que la distribución de nuestros datos no sea tal que los clusters resultantes no sean irrelevantes para nuestros objetivos. Por último, se usó un escalamiento sencillo, utilizando los puntos máximos y mínimos de cada atributo para normalizarlos.

Para el caso del modelo de predicción de churn rate, hicimos uso del pipeline para la transformación de datos que se usó para el modelo no supervisado. Sin embargo, omitimos el tratamiento tan severo de outliers, pues estos siguen dando información relevante para la predicción. Adicionalmente, se hizo un escalamiento robusto, el cual toma en cuenta los rangos intercuartiles para un tratamiento menso radical.

Modelos

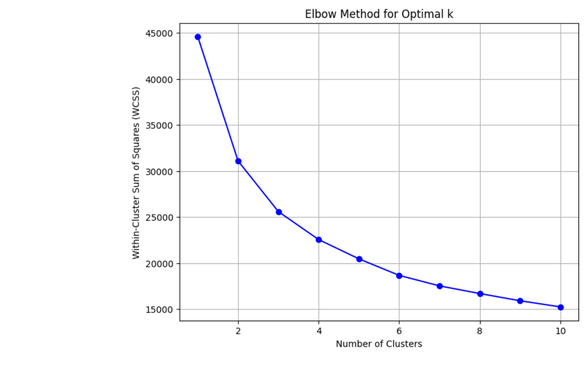
Contamos con varias propuestas para el modelo de clustering. DB-SCAN, por ejemplo, fue uno de los primeros modelos que probamos. Sin embargo, Debido a la distribución tan balanceada de nuestro dataset, la densidad no definía con claridad cada cluster.

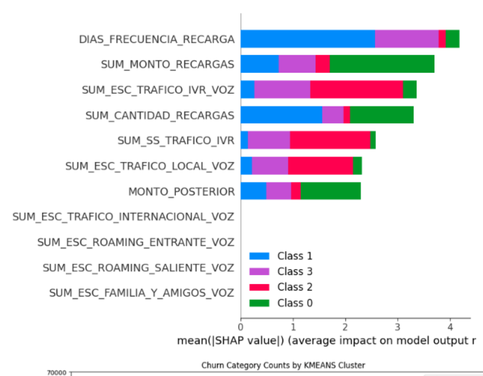
K-Means, sin embargo, pudo obtener clusters de mayor valor, debido al tratado de outliers que manejamos y las distribuciones de datos con las que contamos. De igual manera, no dejaba a clientes sin cluster, algo que era importante para nosotros.

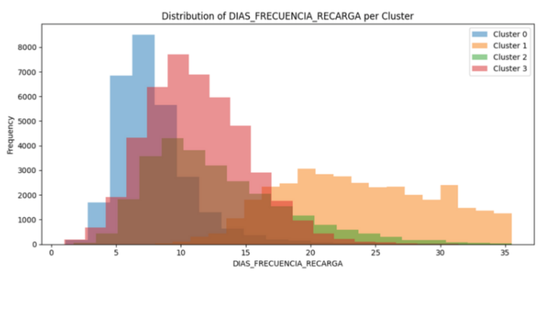
Probamos dos modelos para predecir el churn rate: uno con PyTorch, que tiene una sola capa oculta de 10 neuronas y logra un 90% de precisión en solo 16 épocas, y otro con Light Gradient-Boosting Machine (LGBM), que, aunque no es de deep learning, alcanza un 95% de precisión, siendo más eficiente en términos de tiempo. Ambos modelos son eficaces, pero optamos por LGBM debido a la limitada cantidad de datos, ya que se ajusta adecuadamente al conjunto de datos, aunque PyTorch ofrece flexibilidad para escalar a conjuntos de datos más grandes.

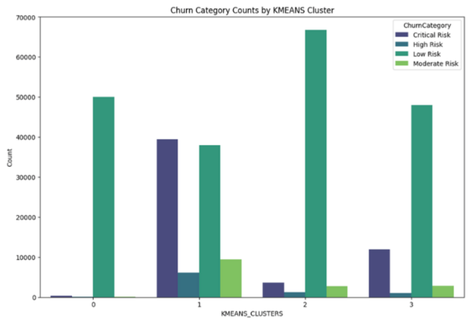
Resultados

Decidimos en cuatro clusters para clasificar a nuestros clientes. Llegamos a este valor utilizando diferentes métricas, como Within-Cluster Sum of Squares para obtener la relevancia de cada cluster, y Shapley Additive exPlanations (SHAP) para obtener la importancia de cada atributo dentro de los diferentes clusters.









Decidimos categorizar nuestros clusters de la siguiente manera:

1. **Recargadores Prodigiosos**
   Usuarios que se caracterizan principalmente por hacer muchas recargas, y de montos grandes. De igual manera, existe muy poco tiempo entre sus recargas.
2. **Recargadores Cautelosos**
   Usuarios caracterizados por muy pocas recargas y de montos pequeños. Cuentan con un tiempo entre recargas muy amplio, en promedio de más de un mes (42 días).
3. **Comunicadores Globales**
   Se puede ver como un subconjunto del cluster 3. Se consideran usuarios "promedio" en cuanto a su uso de recargas. Sin embargo, este cluster se distingue por usar mensajes internacionales. Cuentan con una gran cantidad de mensajes internacionales. Sin embargo, en su mayoría tienen minutos de llamadas nacionales. De igual manera, cuentan con muchos mensajes de voz en su buzón.
4. **Usuarios Estándar**
   Los usuarios promedio. Recargan en promedio cada diez días, muy rara vez usan los servicios de mensajería internacional y buzón de voz. Hacen recargas cuando lo necesita y no gastan más de lo que deben.

Al utilizar LightGBM (LGBM), calculamos el riesgo de Churn por usuario y aplicamos filtros para mostrar estadísticas y gráficos. Categorizamos los riesgos para una mejor comprensión y, mediante ingeniería de datos, clusterizamos a los usuarios por regiones, calculando el riesgo de Churn para cada cluster mediante k-means. Esto proporciona mayor precisión en el marketing dirigido y facilita la comprensión efectiva de los usuarios.

Conclusiones

En este proyecto, abordamos el desafío de analizar datos complejos y variables, destacando la importancia del proceso de análisis de datos en la resolución de problemas. A pesar de las peculiaridades del dataset, logramos implementar estrategias de marketing basadas en clusters regionales.
En cuanto a los objetivos futuros, proponemos ampliar la plataforma con módulos de análisis dinámicos, ofreciendo diversas opciones de clusterización y mejorando la precisión mediante el uso de un LLM más avanzado, como LLaMA. Aunque enfrentamos desafíos, estamos orgullosos de aplicar nuestros conocimientos y tomar decisiones autónomas, estableciendo una base sólida para futuras expansiones y mejoras.