



UNIVERSIDAD DE ESPECIALIDADES ESPÍRITU SANTO
ESAI BUSINESS SCHOOL

SEGUNDO AVANCE

**ANÁLISIS DE LAS VARIABLES MÁS RELEVANTES EN LA TOMA DE DECISIONES
PARA EMPRESA DE TELECOMUNICACIONES**

AUTOR:

WILSON ALEJANDRO DÍAZ GARCÍA

MAESTRÍA:

INTELIGENCIA DE NEGOCIOS Y CIENCIA DE DATOS

ASIGNATURA:

PROYECTO INTEGRADOR DE INTELIGENCIA DE NEGOCIOS

DOCENTE:

ALEXANDRA JACQUELINE ARCINIEGAS CORAL. MGTR.

SAMBORONDÓN

JUNIO 2023

Tabla de contenido

1. Resumen Ejecutivo	3
2. Resumen Ejecutivo Gráfico	3
3. Detalle Técnico	5
3.1 Fuente de datos	5
3.2 Técnicas	5
4. Resultados	6
5. Referencias bibliográficas:	7

1. Resumen Ejecutivo

La empresa de telecomunicaciones enfrenta dificultades para retener a sus clientes debido a la falta de estrategias efectivas. Los usuarios expresan su deseo de cancelar el servicio, lo que representa un gran desafío para la compañía.

El proyecto busca implementar un análisis y sistema de predicción de clientes que deseen cancelar el servicio, esto permitirá mejorar la capacidad de retención y reducir la tasa de abandono. Se utilizarán soluciones basadas en Inteligencia de Negocios y Ciencia de Datos para identificar las variables claves en la toma de decisiones.

La implementación exitosa del proyecto permitirá a la empresa retener a más clientes y reducir la cancelación de servicios. Esto se traducirá en una mejora en la satisfacción del cliente y un aumento en los ingresos de la compañía. Además, la empresa se destacará en el mercado altamente competitivo de las telecomunicaciones y mejorará su posición en comparación con sus competidores. La solución brindará una ventaja estratégica al proporcionar información valiosa para la toma de decisiones basada las variables más importantes o relevantes.

2. Resumen Ejecutivo Gráfico

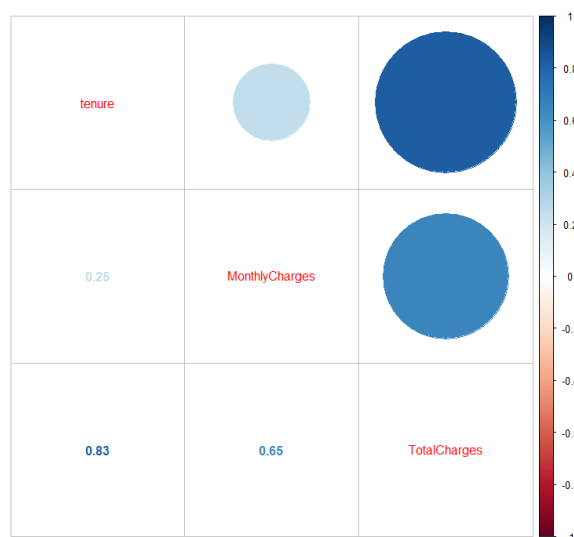


Figura 1. Correlación de variables numéricas: Tenure, MonthlyCharges y TotalCharges

En el gráfico, figura 1, se observan las correlaciones entre las variables numéricas Tenure, MonthlyCharges y TotalCharges.

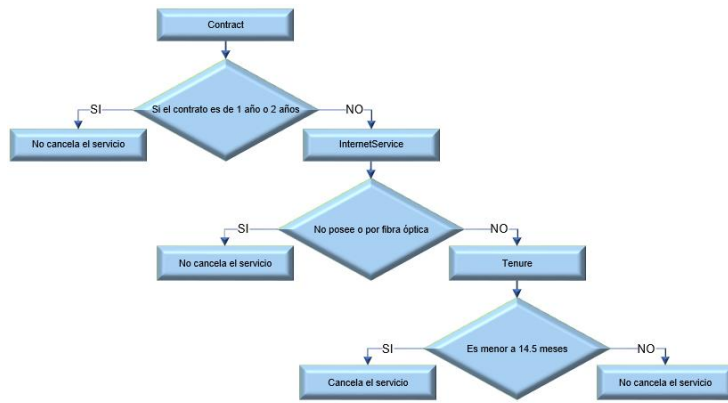


Figura 2. Árbol de decisión para determinar si el cliente desea o no la cancelación del servicio

En el gráfico presentado, figura 2, se puede apreciar que las variables que son más importantes en la toma de decisiones según el árbol de decisión podado.

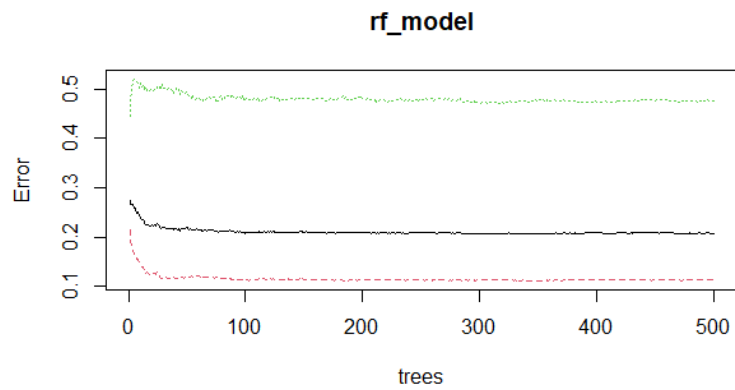


Figura 3. Random Forest Model

En el gráfico presentado, figura 3, se aprecia la estabilización de los árboles generados de manera independiente.

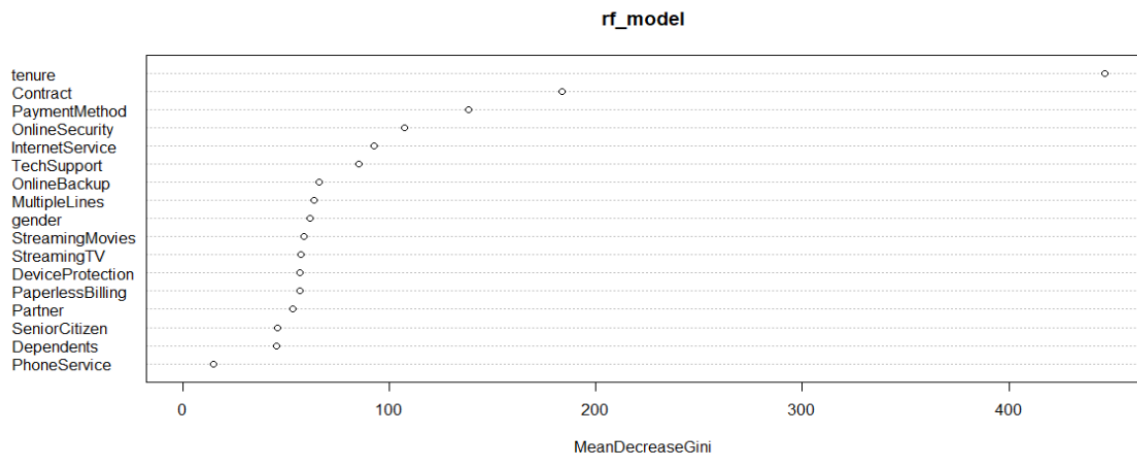


Figura 4. Importancia de variables

En el gráfico presentado, figura 4, se observan las variables más importantes para el modelo Random Forest (James, Witten, Hastie, & Tibshirani, 2013), entre ellas las 3 principales son: tenure, Contract y PaymentMethod.

3. Detalle Técnico

3.1 Fuente de datos

La fuente de datos fue obtenida de Kaggle, cuenta con una base de datos de 7.043 usuarios, las variables a utilizar son: género, si el usuario es mayor a 65 años, si posee pareja, si vive con dependientes, el número de meses que tiene el servicio con la empresa, si posee servicio de telefonía, si posee varias líneas telefónicas, si posee servicio de internet con la empresa, o el servicio que tiene, si posee servicio de seguridad online, si posee un servicio de respaldo de datos, si posee servicio de detección de dispositivos, si posee plan de soporte técnico, si posee servicio de tv, si posee servicio de películas, el tipo de contrato, si tiene servicio de facturación electrónica, el modo de pago, el valor mensual a pagar por servicios, el valor total del cliente y si el cliente desea o no cancelar el servicio; con la data en mención se busca identificar las variables más relevantes en la predicción de la cancelación del servicio por parte del cliente, con el fin de desarrollar estrategias que permitan mejorar la calidad del servicio y, por ende, fidelizar a los usuarios.

Para trabajar los algoritmos de Decision Tree y Random Forest no se tomaron en cuenta las variables customer id, cargo mensual y cargo total, la primera es irrelevante y las dos últimas son valores de cobro dependientes de los varios servicios que posee el abonado.

3.2 Técnicas

Previo al uso de las técnicas se valida la correlación entre las variables numéricas (figura 1), en las que se observa que las variables que poseen una fuerte correlación son tenure y Monthlycharges vs Totalcharges, poseen 0.83 y 0.65 respectivamente, poseen cierta dependencia lineal o relación entre ellas.

Las variables numéricas que posee menor correlación son tenure con Monthlycharge, es 0.25, hay baja dependencia lineal o poca relación entre ellas (Hastie, T., Tibshirani, R., & Friedman, J. (2009)).

Las técnicas usadas para predecir las variables más importantes son:

Árbol de decisión y podado, (James, Witten, Hastie, & Tibshirani, 2013), es una herramienta de análisis que permite interpretar y explicar fácilmente las decisiones tomadas, está orientado a identificar los factores que inciden en la decisión del usuario. Además, muestran la relevancia de las variables en la predicción, y la técnica de podado elimina las ramas del árbol que no aportan significativamente a la precisión de la predicción.

Random Forest, (James, Witten, Hastie, & Tibshirani, 2013), "es un método de aprendizaje en conjunto versátil y potente que combina las predicciones de múltiples árboles de decisión para ofrecer resultados robustos y precisos en una amplia variedad de tareas de predicción", en el gráfico, figura 3, se puede apreciar cómo se estabiliza los múltiples arboles generados de manera independiente, utilizando una partición recursiva que busca la mejor división de los datos en cada nodo del árbol, se observa que se estabiliza en los 20 árboles aproximadamente, los datos del usuario se representan con colores: verde para los que no cancelan el servicio, rojo para los que sí cancelan y negro para los generados pero que no visibles, pero están dentro de la corrida estadística.

4. Resultados

En base a los resultados del decision tree (figura 2), se puede concluir que las variables que poseen mayor peso en la toma de decisiones son: Contract, InternetService y tenure, por ello es necesario implementar tácticas de mejora que permitan optimizar el servicio ofrecido y así garantizar la satisfacción y retención de los clientes a largo plazo.

Según los resultados del Random Forest (figura 4), se puede observar que las variables que poseen mayor importancia en la toma de decisiones son tenure, contract y paymentmethod, respectivamente.

Al comparar las dos técnicas de predicción, se puede observar que Random forest obtiene una tasa de error del 20.79%, mientras que la técnica de árbol de decisión presenta una tasa de error ligeramente mayor del 21.50%, lo que sugiere que Random forest es más precisa para predecir los resultados (Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.).

5. Referencias bibliográficas:

- Lampubhutia, L. (s. f.). Telecom-Customer-Churn. Kaggle. Recuperado el 16 de Junio de 2023, de <https://www.kaggle.com/datasets/lampubhutia/telecomcustomer-churn>
- Statinfer. (s. f.). Pruning a Decision Tree in R. Recuperado el 16 de Junio de 2023, de <https://statinfer.com/203-3-10-pruning-a-decision-tree-in-r/>
- R-Tools Technology. (08 de Abril de 2018). R-STUDIO. Recuperado 16 de Junio de 2023, de R-STUDIO: <https://www.r-studio.com/index.shtml>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). An Introduction to Statistical Learning: With Applications in R (Springer Texts in Statistics) (Corrected at 6th printing 2017 edition). Springer. Recuperado 16 de Junio de 2023
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. Recuperado 16 de Junio de 2023
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media. Recuperado 16 de Junio de 2023