**Universidad de Costa Rica**

**SP-6532 Econometría Avanzada- Proyecto Final**

**Valor 30%**

**Fecha límite de entrega: 7/12/2022.**

Buenas tardes estimada profesora, espero que se encuentre bien. Para mí es un gusto presentarle este proyecto de investigación en donde pongo a su conocimiento de qué trata la investigación y los principales resultados.

Este investigación pretende resolver qué variables están contribuyendo en la fuga de los clientes de la subscripción de compañías de telecomunicaciones, cuáles características de los clientes hacen más propensos que los clientes hagan fuga, y finalmente, desea resolver qué acciones se pueden tomar por parte de las empresas de telecomunicaciones para evitar la fuga.

Con el desarrollo vertiginoso de la industria de las telecomunicaciones, los provedores de servicios cada vez están más inclinados hacia la expansión de su base de clientes leales. Para mantenerse a flote dentro del clima competitivo, retener consumidores existentes se ha vuelto el principal reto: según la literatura, el costo de atraer a un nuevo cliente es mucho mayor que retener a uno contento, es por esto, que es necesario que las empresas de telecomunicaciones utilicen avanzadas técnicas de análisis de datos para entender el comportamiento de sus clientes y predecir el nivel de certeza de que un cliente deje o no la compañía.

Esa base de datos contiene información de los consumidores para una compañía. Varios atributos sobre los servicios usados son recopilados en la base.

Es importante determinar cuáles son las variables que contiene la base de datos y su significado:

Churn

1 if customer cancelled service, 0 if not

AccountWeeks

number of weeks customer has had active account

ContractRenewal

1 if customer recently renewed contract, 0 if not

DataPlan

1 if customer has data plan, 0 if not

DataUsage

gigabytes of monthly data usage

CustServCalls

number of calls into customer service

DayMins

average daytime minutes per month

DayCalls

average number of daytime calls

MonthlyCharge

average monthly bill

OverageFee

largest overage fee in last 12 months

RoamMins

average number of roaming minutes

La pregunta que se desea responder es en qué medida, las variables características de una persona usuaria aumentan la probabilidad de clasificar a esa persona como una desertora de los servicios de comunicación que les brinda una empresa.

Primeramente, es importante notar en la variable CHURN, la frecuencia de personas que abandonaron la suscripción respecto a quienes no. Esta información muestra un desbalance importante entre ocurrencias. Esta situación contribuye a un problema de clasificación sobre la proporción de la variable respuesta en datos\_train y datos\_test cuando se realiza la validación cruzada.

Con respecto a la variable ContractRenewal se espera que quienes hayan renovado el contrato telefónico aporten en la probabilidad de quedarse en la compañía actual, pues habiendo no podido renovar, deciden hacerlo, pues poseen un grado de preferencia por la empresa. Para aquellos que no renuevan, lo subsiguiente es que salgan del contrato.

Finalmente, con la variable DataPlan sobre si la persona posee un plan de datos, se espera que, algún tipo de plan de datos sea congruente con la salida de la persona de la suscripción. A nivel de la forma en que trabajan estos modelos de negocio, se espera que aquellas personas que poseen plan de datos, sean aquellas que cambien de suscripción en el tanto ellas esperaban algún nivel mínimo de calidad de los datos y no se vieron satisfechan, por lo que serán más propensas a abandonar.

AccountWeeks muestra la cantidad de semanas en que la persona ha poseído una cuenta activa, una posible transmisión de variables mostraría que valores muy bajos representa una mayor ocurrencia de abandono de la suscripción pues implica un descontento directo.

DataUsage muestran los gigabytes mensuales de uso de datos, esta variable continua posee una gran cantidad de valores en 0 a 0.54, más de la mitad, es por esto, que se propone convertir esta variable en dicotómica de 0 a 0.54, y de 0.55 en adelante de manera que se pueda analizar un patrón de si gigabytes bajos de datos implica que la persona no le da uso a la subscripción y es por esa razón que la abandona.

CustServCalls muestra la cantidad de llamadas de servicio al cliente que ha hecho la persona usuaria. Se espera que mayores llamadas, aumenten la probabilidad de clasificarse como una persona que abandona la subscripción. La cantidad de llamadas del los primeros 3 cuantiles es de 2, pero el máximo de llamadas es de 9, por lo tanto, se espera que valores extremos por arriba de 2, representen mayor propensión del abandono.

DayMins y DayCalls muestras la cantidad de minutos y de minutos de llamada que usa la persona en la subscripción, se espera que un aumento de la cantidad de minutos, clasifique a la persona como alguien que se mantiene en la subscripción. Minutos bajos pueden indicar el poco uso de la subscripción, por lo que promueve el desinterés por la subscripción.

MonthlyCharge muestra el costo promedio mensual de la subscripción. Esta variable sufre un problema de endogeneidad, el costo de la subscripción puede estar correlacionada con la cantidad de minutos de datos, llamadas y de roaming, sin embargo, la base de datos no lo especifica. Para nuestros fines, la vamos a tratar como una variable independiente, de manera que se espera que un aumento en el costo pueda significar un aumento de la clasificación a abandono.

OverageFee representa el mayor cargo por excedente en los últimos 12 meses. En algunas ocasiones, cuando cae de sorpresa un cargo por excedente, las personas tienden a asustarse y a tomar control de la situación antes que sus finanzas le limiten, es por eso que optan por cancelar inmediatamente la subscripcción en estos escenarios. Se espera que, a raíz de cargos por excedentes altos, estos cancelen la subscripción.

Finalmente se tiene RoamMin, este mide el número de minutos en estado de roaming que usa el usuario.

Sobre el tratamiento y preparación de datos, se nota que ninguna variable posee valores nulos y la única transformación consistente de datos es la de la variable DataUsage, de manera que provea información más valiosa de ella, tal como se explicó sobre la hipótesis de transmisión de variables de la etapa anterior.

A partir de la revisión de datos atípicos mediante boxplots utilizando un criterio de rango intercuantílico.

Los resultados muestran que de 3333 observaciones para cada variable, en su mayoría, los valores atípicos no sobrepasan a las 20 observaciones a excepción de RoamMins y CustServCalls. A este último, al representar la cantidad de llamadas a servicio al cliente, es mejor no hacer ninguna transformación en el tanto muchas llamadas pueden ser parte de la queja sostenida del cliente. A RoamMins se le hará caso omiso pues igualmente siguen siendo pocos valores y es más preciso realizar una buena clasificación que ajustar mejor el nivel de la media.

Finalmente, para la etapa de modelamiento, usamos dos modelos de clasificación: el modelo logístico y el K vecinos más cercanos. Cuando incluimos todas las variables importantes que deseamos considerar como influyentes en el modelo.

Primeramente, podemos ver el modelo logístico, ver qué factores afectan positivamente a la elección de la persona promedio y reconocemos la calidad de la predicción.

En un inicio, se puede ver el nivel de desbalance en la frecuencia de la variable de clasificación, con un 14.5% de renuncia.

Una vez que corremos el modelo y lo testeamos, encontramos que el accuracy tiene un buen nivel, sin embargo, el Kappa es bajo, esto es debido al desbalance existente en la variable de clasificación. Accuracy es de 86%, pero Kappa es sólo de 21.7%.

Cuando revisamos los coeficientes, poseemos lo siguiente:

Los coeficientes de un modelo Logit no se pueden interpretar, pero la dirección sí: note que los coeficientes significantes poseen una relación positiva con el abandono de la subscripción.

Conforme la muestra en promedio nunca ha renovado el contrato, aumenta la probabilidad de renunciar a la subscripcción.

Conforme aumentan las llamadas de servicio al cliente, se aumenta la probabilidad de renunciar a la subscripción.

Si aumentan los minutos de Roaming, por alguna razón, se renuncia a la subscripción, esto puede estar relacionado a que el tipo de subscripción no es el adecuado para la persona.

Cuando vemos la matriz de confusión se puede notar el gran desbalance (también notable en Kappa) entre los valores de renuncia de la subscripción respecto a la estancia en la subscripción. Esto repercute en la facilidad del modelo por predecir las observaciones que sí renuncian a la subscripción. Los valores más altos son la especificidad y la asertividad negativa.

Al lanzar el modelo de K vecinos más cercanos, se encuentra información interesante, no necesariamente una estimación, pero sí una predicción más apropiada que del modelo logístico.

El modelo indica que la cantidad de vecinos K óptima es de 5, pues asegura el menor accuracy y kappa.

Cuando se aplica el modelo para predecir nuevos datos, este modelo predice mejor que el logístico. Tal como menciona la teoría respecto a los desbalances de la clasificación, como en el ejemplo que poseemos, el Kappa muestra la calidad real del modelo pues normaliza respecto a la calidad de predicción de los positivos. El modelo es muy bueno en Specificity: 0.98315, Pos Pred Value : 0.83099, Neg Pred Value : 0.91984. Esto indica que el modelo es muy bueno clasificando a las personas que no abandonan.

En conclusión, para esta investigación nos quedamos con el modelo de k vecinos más cercanos. Si parece relevante, tal como lo mencionan algunos críticos de la base de datos, revisar bien si las variables usadas son un buen proxy de la realidad de la clasificación.