Ejercicio Final
Curso Superior en Ciencia de Datos y Big Data - MasterD

Introducción

En este ejercicio final vamos a trabajar con un conjunto de datos alojado en la plataforma Kaggle. Kaggle es una comunidad online de científicos de datos propiedad de Google. Kaggle permite a los usuarios encontrar y publicar conjuntos de datos, explorar y construir modelos, así como colaborar con otros científicos de datos y participar en competiciones en la que se plantean desafíos de análisis de datos.

En primer lugar, entraremos en la plataforma de Kaggle, en la sección de datasets para descargar los datos relacionados con el ejercicio. Accede al siguiente enlace

https://www.kaggle.com/blastchar/telco-customer-churn

y descarga los datos. Para poder descargar los datos, primero deberéis registraros antes en la plataforma Kaggle.

Una vez descargados los datos, tendréis un fichero zip que al descomprimirlo obtendréis un fichero csv, con los datos con los que vais a trabajar.

El conjunto de datos con el que vamos a trabajar contiene datos de una muestra de clientes de una compañía de telecomunicaciones.

El objetivo del proyecto de ciencia de datos que vais a realizar es predecir por un lado, el gasto de los clientes y por otro lado predecir los clientes en riesgo de cambiar de compañía (se suele denominar "*churn*" en inglés) en el mes en curso y realizar un modelo predictivo y una simulación de una campaña para retener a los clientes.

Cada fila del fichero representa un cliente e incluye información sobre:

- Clientes que se fueron en el último mes: la columna se llama Churn
- Servicios que cada cliente tiene contratados: teléfono, varias líneas, internet, seguridad, soporte técnico, protección de dispositivos, o servicio de TV y películas en streaming.
- Información de la cuenta del cliente: antigüedad en la compañía, duración del contrato, método de pago, facturación electrónica, cargos mensuales y cargos totales
- Información demográfica sobre los clientes: género, rango de edad y si tienen pareja o dependientes a su cargo

A continuación, se detalla el significado preciso de las columnas del conjunto de datos

- customerID: Identificador de cliente
- gender: Varón(male) o mujer (female)
- SeniorCitizen: El cliente es de la tercera edad o no (1,0)
- Partner: El cliente tiene pareja o no (Yes,No)
- Dependents: El cliente tiene dependientes a su cargo o no (Yes,No)

- Tenure: Antigüedad en la compañía en meses
- PhoneService: El cliente tiene contratado servicio de teléfono o no (Yes, No)
- MultipleLines: El cliente tiene contratadas múltiples líneas (Yes, No, No phone service)
- InternetService: Tecnología de Internet contratada (DSL, Fiber optic, No)
- OnlineSecurity: El cliente tiene contratado el servicio de seguridad online (Yes, No, No internet service)
- OnlineBackup: El cliente tiene contratado servicio de backup (Yes, No, No internet service)
- DeviceProtection: El cliente tiene contratado el servicio de protección de dispositivos (Yes, No, No internet service)
- TechSupport: El cliente tiene contratado el servicio de soporte técnico (Yes, No, No internet service)
- StreamingTV: El cliente tiene contratado el servicio de TV en streaming (Yes, No, No internet service)
- StreamingMovies: El cliente tiene contratado el servicio de películas en streaming (Yes, No, No internet service)

- Contract: Duración del contrato (Month-to-month, One year, Two year)
- PaperlessBilling: El cliente recibe facture electrónica (Yes, No)
- PaymentMethod: Modo de pago (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
- MonthlyCharges: Tarifa mensual
- TotalCharges: Cargos totales facturados al cliente en todo el periodo de antigüedad
- Churn: El cliente ha abandonado la compañía en el mes en curso o no (Yes or No)

Objetivos y Entregables

Los objetivos del proyecto son los siguientes:

- Realizar un análisis exploratorio completo del conjunto de datos para poder descubrir anomalías en los datos y descubrir patrones que relacionen a las distintas variables
- Modelos de regresión: Diseñaremos y ajustaremos modelos de regresión lineal para predecir el gasto mensual de los clientes (MonthlyCharges) a partir de los servicios contratados y sus características personales
- Modelos de clasificación para predecir el abandono (Churn) de los clientes
- Evaluación de campañas de retención a partir de los resultados de los modelos de Churn.

El entregable será un fichero de Rmarkdown o RNotebook .Rmd y el fichero html que se genera al compilarlo (Knit) en Rstudio. El fichero .Rmd debe contener todo el código en R para producir los resultados que se especificarán a continuación. El fichero .Rmd debe poder ser compilado por el profesor sin errores (se permiten warnings).

Se proporciona una plantilla de fichero (plantilla_ejercicio_final.Rmd) con el esqueleto principal del documento para que no partáis de cero.

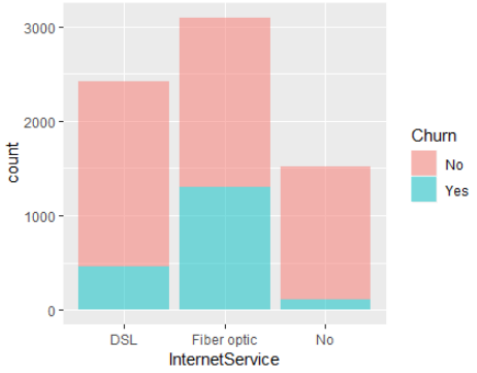
Desarrollo del Ejercicio

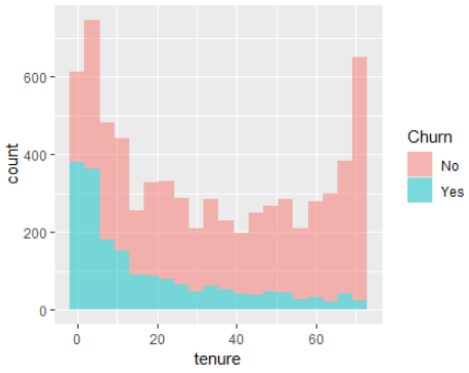
En primer lugar, cargaremos los datos en R, leyendo el fichero csv. Almacenamos el contenido del fichero en dataframe llamado churn_data

Análisis Exploratorio

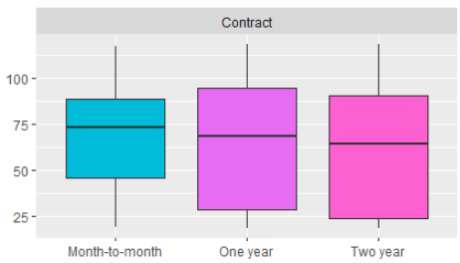
Lo realizaremos con el conjunto completo de datos

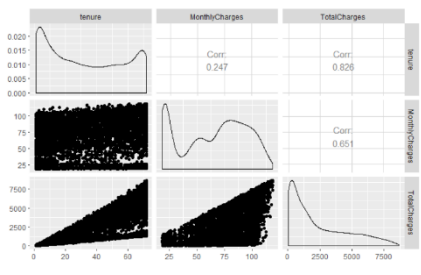
- Identifica para columna el tipo de dato (character, factor, numeric, etc) en el que ha sido almacenado
- Visualiza un resumen del conjunto de datos mediante la función summary.
- De cara a la predicción del churn, para cada una de las variables excepto el customerID:
    - Realiza un diagrama de frecuencias para las variables cualitativas o histogramas para las variables cuantitativas
    - Muestra los gráficos anteriores coloreados según la variable Churn para poder ver la proporción de "churners" en función del valor de las distintas variables. Se muestran dos ejemplos a continuación





- De cara a la predicción del gasto mensual:
    - Para las variables categóricas realiza un boxplot que muestre la distribución del gasto mensual (MonthlyCharges) para cada valor de las variables
    - Para las variables numéricas realiza un gráfico de dispersión que muestre la correlación entre variables. Puedes hacerlo uno a uno o usar un gráfico de correlación conjunto.
      Se muestran dos ejemplos a continuación





- Cualquier otra cosa que se te ocurra y creas que da información sobre los problemas que queremos resolver en el proyecto.

En este proyecto en particular, los datos están limpios y no hay valores erróneos o anómalos. En general, esto es extraño que suceda y el proceso de análisis exploratorio te hubiera ayudado a encontrar y corregir estos errores.

Si hay alguna variable que deba ser considerada como categórica en lugar de como numérica transfórmala. Es necesario para que los modelos de regresión o clasificación funcionen correctamente.

Antes de seguir adelante, vamos a dividir el conjunto de datos en dos subconjuntos de entrenamiento y test:

- churn_train: Contendrá las 5000 primeras de churn_data
- churn_test: Contendrá las filas de la 5001 al final de churn_data

Modelos de Regresión

Vamos a diseñar y ajustar un modelo de regresión para modelar y predecir el gasto mensual de los clientes

Ajusta un modelo de regresión multilineal sobre los datos de entrenamiento churn_train:

- Usa como variable objetivo MonthlyCharges
- Usa como variables predictoras todas las demás menos customerId, TotalCharges y Churn.

Muestra los detalles del modelo ajustado mediante la función summary aplicada al objeto donde has almacenado el modelo ajustado. A partir del detalle sobre los coeficientes del modelo ajustado responde a las siguientes cuestiones:

- ¿Cuáles son las variables/coeficientes que a un nivel de significancia del 95% podemos concluir que son no nulos?
- A partir del valor de los coeficientes, ¿podrías estimar el coste de los diferentes servicios: teléfono, DSL, Fibra, seguridad, TV, etc.?

Ahora a partir del modelo ajustado, haz una predicción sobre el conjunto de test: churn_test y calcula el error promedio de las predicciones. Como métrica del error del las predicciones usaremos el RMSE (root mean square error) : $RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2}$

Modelos de Clasificación

Vamos a diseñar y ajustar un modelo de Clasificación para predecir los clientes que pueden abandonar la compañía.

Ajusta un modelo de regresión logística sobre los datos de entrenamiento churn_train:

- Usa como variable objetivo Churn
- Usa como variables predictoras todas menos customerId.

Con el modelo ajustado, predice sobre los datos de test (churn_test) la probabilidad de abandono de los clientes. A partir de dicha probabilidad asigna a una variable churn_pred =”Yes” si prob >0.5 y “No” en caso contrario. Calcula la matriz de confusión de la predicción:

	Pred	
Real	**No**	**Yes**
No	TN	FP
Yes	FN	TP

A partir de ella calcula:

- La precisión global de la predicción: Accuracy = (TN + TP) / (TN + FN + FP + TP)
- La ratio de falsos positivos: FPR = FP / (FP + TN)
- La ratio de falsos negativos: FNR = FN / (FN + TP)
- La ratio de verdaderos positivos: TPR = TP / (FN + TP)

Ahora, ajusta en el conjunto de entrenamiento un modelo de clasificación para la variable Churn que use como predictores solo Contract, Tenure e InternetService.

Predice en el conjunto de test y calcula la matriz de confusión, así como las métricas que has calculado para el modelo anterior y compara los resultados.

Simulación de Campaña de Retención

Ahora a partir de los resultados del primer modelo de regresión logística (en el que usabas todas las variables) vamos a evaluar el rendimiento económico de una campaña de retención. Usaremos las predicciones sobre el conjunto de test.

Para calcular el rendimiento de la campaña tendremos en cuenta lo siguiente:

- A los clientes clasificados por el modelo como “Churners” se les ofrecerá un teléfono de regalo a cambio de que firmen un contrato de permanencia por un año. El coste del teléfono lo denominamos CT. Estos clientes aceptaran el trato con una probabilidad AR.
- El beneficio asociado a la retención de un cliente lo estimamos como el consumo medio anual de un cliente menos el coste de gestión (uso de infraestructura, facturación, atención al cliente etc). Estimaremos que el retorno medio de la retención de un cliente es R (lo fijaremos en 500 Euros).
- Un cliente que abandona la compañía, estimamos que genera una pérdida igual al retorno R que genera la retención, es decir 500 Euros

Para diferentes umbrales de clasificación up = (0, 0.1,0.2,0.3, 0.4, 0.5, 0.6 ,0.7, 0.8, 0.9, y 1), en función de la probabilidad de Churn estimada por el modelo clasificaremos a los clientes en pred_churn = ”Yes” si prob>up y “No” en caso contrario. En el apartado anterior hemos utilizado la elección habitual de up=0.5.

Conforme suba el umbral de probabilidad, seleccionaremos a menos clientes para la campaña. Por un lado, esto tendrá menos coste debido al incentivo que se ofrece, pero por otro lado se obtendrán menos retenciones. Hay que encontrar el umbral óptimo en el que se obtiene un mayor rendimiento para la campaña.

Para cada umbral de probabilidad, calculamos a partir de la probabilidad del modelo la variable churn_pred y calculamos la matriz de confusión

	Pred	
Real	**No**	**Yes**
No	TN	FP
Yes	FN	TP

Podemos estimar el beneficio de la campaña, para ese umbral, como

Beneficio = ResultadoConCampaña – ResultadoSinCampaña

y

ResultadoConCampaña= - FP*AR*I - TP*AR* I - TP*(1-AR)*R - FN*R

donde

- El primer y segundo término corresponde el gasto en los teléfonos que regalamos a los clientes que aceptan la promoción y que no hubieran abandonado la compañía, aunque no se hubiera hecho campaña.

- El tercero representa la perdida de los clientes a los que se ha hecho la promoción, no la han aceptado y se han marchado.

- El cuarto representa a los clientes a los que no se ha realizado la promoción y han acabado marchándose.

ResultadoSinCampaña = - (FN + TP) *R

- Este valor representa la pérdida de ingresos debido al abandono real de los clientes, en caso de no hacerse campaña. Notad que ese término no depende del modelo ajustado ni de los umbrales de probabilidad.

Nota que el resultado en ambos casos es un número negativo. Solo tenemos en cuenta pérdidas económicas, ya sea por gasto de promoción o por la facturación esperada a futuro de un cliente que abandona la compañía.

Vamos a calcular los resultados de la campaña para dos escenarios de incentivo distintos.

1) Supongamos:
    - Coste teléfono de regalo: I=200 Eur
    - Probabilidad de aceptación AR= 0.4
    - R = 500 Eur

Calcula para cada umbral de probabilidad, el beneficio de la campaña usando las fórmulas anteriores. ¿Cuál sería el umbral de probabilidad para la selección de clientes óptimo para la campaña?. ¿Qué beneficios generaría?

2) Supongamos:
    - Coste teléfono: I=400 Eur
    - Probabilidad de aceptación AR= 0.8
    - R = 500 Eur

Calcula el umbral óptimo y el beneficio de la campaña para el umbral óptimo. ¿Se obtienen mejores o peores resultados respecto a la campaña con incentivo menor?