Universidad de Montevideo – Introducción a la Ciencia de Datos

Parcial 2020

14 de Octubre, 2020

1. Suponga que una empresa de telefonía móvil lo contrata para entender por qué sus clientes lo están abandonando. Como insumo, usted obtiene datos de consumo, número de quejas, retrasos en el pago de servicios, antigüedad del cliente y tipo de contrato.

    a) Diagrame la estructura o workflow del proyecto con sus principales características y detalle los documentos que van a ser entregados al final del proyecto.
    b) Como clasificaría las siguientes preguntas de acuerdo a los *tipos de preguntas* vistos en clase, explique su respuesta:
        - En base a una muestra representativa de clientes cuáles son los principales factores que causan el abandono.
        - Pronosticar cuál es la probabilidad de que un cliente abandone la empresa.

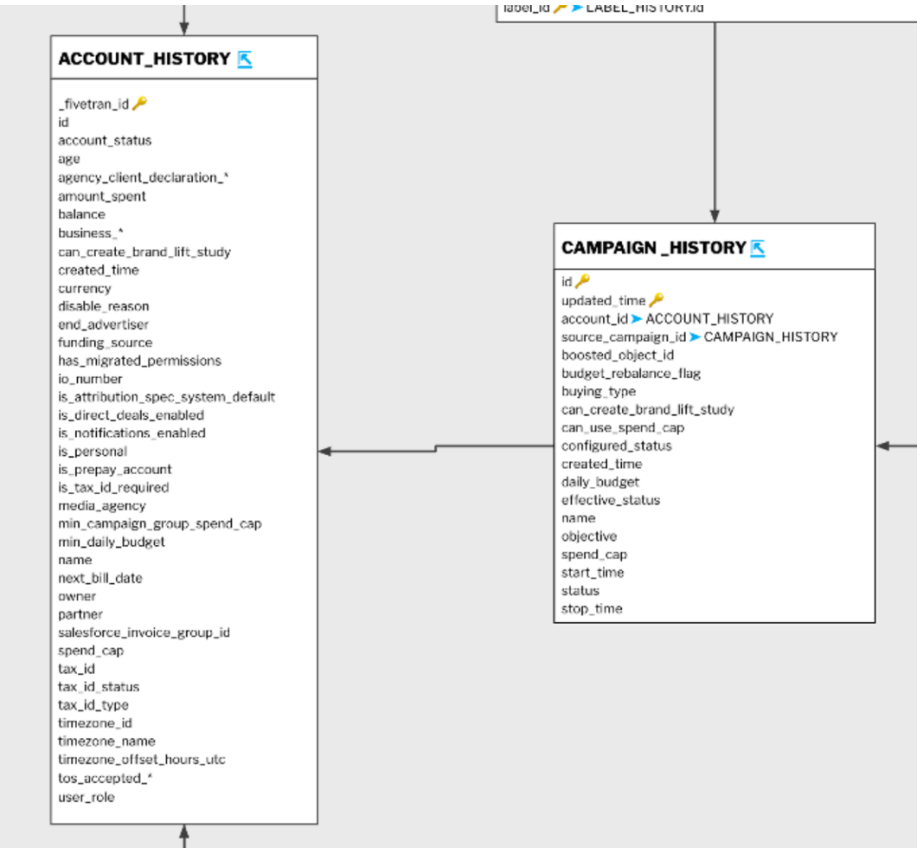
2. Se le encarga el trabajo de hacer un modelo para predecir el precio de propiedades inmobiliarias y se le entrega el siguiente dataset de variables explicativas:

First Floor Sq Ft	Type	Kitchen Quality	Neighborhood
725	60	Gd	Sawyer
913	60	Gd	SawyerW
1057	20	Gd	NAmes
744	60	TA	Timber
831	50	TA	SawyerW
1888	20	Gd	NAmes
1072	180	TA	Edwards
1188	20	TA	NAmes
924	20	TA	OldTown
1040	60	Fa	NAmes

First Floor Sq Ft se refiere al tamaño de la propiedad medido en pies. *Type* hace referencia al tipo de apartamento, que surge de un diccionario que te entrega el cliente, se utilizan números por razones prácticas. *Kitchen Quality* refiere a la calidad de la cocina (Ex = Excellent, Gd = Good, TA = Typical/Average, Fa = Fair, Po = Poor). *Neighborhood* indica el nombre del barrio donde se encuentra la propiedad.

    a) Clasifique las variables de acuerdo al tipo de datos que son y argumente su elección.
    b) ¿Cómo considera que es apropiado incorporar la variable *Type* en un modelo de regresión lineal? Explique su respuesta.

3. Usted dispone del siguiente diagrama entidad relación (DER). El mismo es un DER real, de una ETL tool llamada FiveTran, la cual su jefe le ha dicho que use para extraer datos de Facebook Ads de sus clientes. De este DER a usted le interesan las dos tablas que se le presentan en la imagen. La tabla ACCOUNT HISTORY contiene datos históricos de las cuentas de los usuarios, mientras que CAMPAIGN HISTORY contiene datos históricos de las campañas publicitarias que se le han mostrado a estos usuarios. Muchas campañas publicitarias pueden serles mostradas a muchos usuarios. Usted decide hacer una query a esta base de datos con el fin de obtener todos los ID y nombres de los usuarios. Esta es la tarea principal de la query. Además, en esta query usted quiere también extraer los ID y nombres de las campañas que le han sido mostradas a los usuarios, si es que tienen. ¿Qué operación debe hacer entre éstas dos tablas y por qué? Indique qué campos utilizaría y por cuales conectaría ambas tablas (key y foreign key).



4. ¿Es bueno guiarse únicamente por macro estadísticas como la media y la desviación estándar para conocer la distribución de un conjunto de datos? Argumente su respuesta y de un ejemplo.

5. ¿Por qué dos variables altamente correlacionadas entre sí pueden llegar a ser buenas para predecir una a la otra sin necesariamente existir causalidad entre ambas?

6. Se sabe que la alcalinidad, en miligramos por litro, del agua en los tramos superiores de los ríos en una región particular se distribuye normal con una desviación estándar de 10 mg/l. También se sabe que las lecturas de alcalinidad en los tramos inferiores de los ríos de la misma región tienen una distribución normal, pero con una desviación estándar de 25 mg/l.

Se realizan diez lecturas de alcalinidad en el curso superior de un río y quince en el curso inferior del mismo río con los siguientes resultados.

Upper reaches	91	75	91	88	94	63	86	77	71	69
Lower reaches	86	95	135	121	68	64	113	108	79	62
	143	108	121	85	97					

Confianza = 90% -> Critical Value Z = 1.2816
Confianza = 95% -> Critical Value Z = 1.6449
Confianza = 99% -> Critical Value Z = 2.3263

Investigar, al nivel de significancia del 1%, la afirmación de que la verdadera alcalinidad media del agua en los tramos inferiores del río es mayor que en los tramos superiores.

a) Plantee la prueba de hipótesis.
b) Calcule el estadístico Z.
c) Responda la pregunta planteada y comente los resultados.

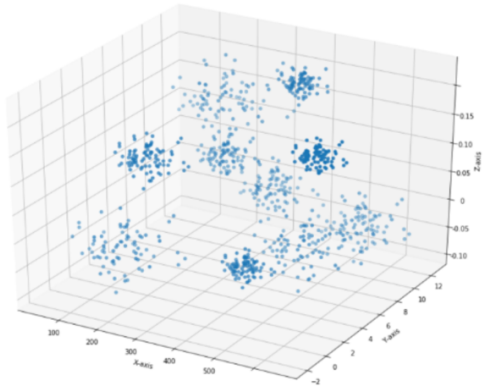
7. Usted y un amigo están comiendo una bolsa de M&M's, cuando su amigo dice: "Parece que hay más caramelos amarillos y marrones que caramelos rojos y azules. De hecho, afirmo que hay un 30% amarillo, 30% marrón, y solo el 20% rojo y 20% azul."

   Juntos cuentan los M&M's en la bolsa y obtienen los resultados a continuación:

Color	Amarillo	Marrón	Rojo	Azul	**Total**
Cantidad	58	61	55	46	**220**

   a) Utilice el método de valor crítico con nivel de significancia 0.05 para probar la afirmación de su amigo. Tenga en cuenta que el valor crítico para una Chi-Cuadrado al 95% con 3 grados de libertad es 7.815.

8. Usted dispone de un dataset de encuestas sobre un producto. Un experto le ha dicho que existen tres preguntas cuyas respuestas determinan casi perfectamente si el encuestado compraría el producto o no. Sin embargo, algunos encuestados no han contestado una de las preguntas porque le ha incomodado. Usted no sabe cuántos encuestados no han contestado esa pregunta, pero sabe que cada encuesta es vital y ha costado mucho dinero. Indique cómo imputaría los valores faltantes, indicando qué pre-procesamiento haría y qué modelo utilizaría.



9. A continuación se presentan tres gráficos. ¿Le parecen apropiados? ¿Qué criticas les haría?

Gráfico 1 - Número de pruebas de COVID-19 en Argentina

Gráfico 2 – Mercado de Smartphones en Estados Unidos

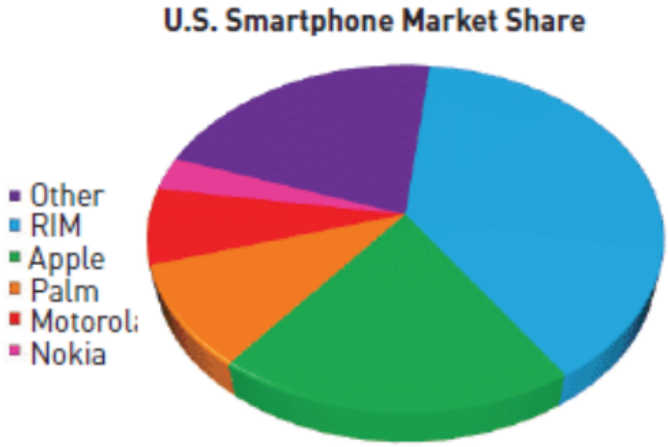
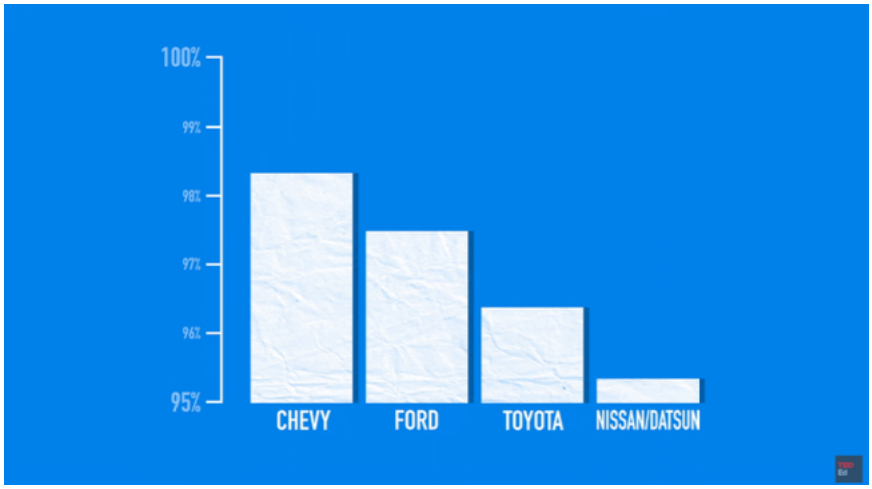


Gráfico 3 - 98 percent of Chevy trucks sold in the past 10 years are still on the road



10. Suponga que está interesado en entender qué variables impactan el consumo de Carne Vacuna en Estados Unidos y en qué magnitud. Para eso, realiza un modelo de regresión lineal simple con el Consumo Anual de Carne Vacuna (Beef) como variable dependiente y el Consumo Anual de Cerdo (Pork) como variable independiente.

    A continuación se presenta el resultado de la regresión:

Average Annual Beef Consumption vs. Average Annual Pork Consumption

Regression Statistics	
Multiple R	0.2915
R^2	0.0850
Adjusted R^2	0.0240
Standard Error	4.2286
Observations	17

F test results	
F	*Signif. F*
1.39	0.2563

	Coefficients	*Std Error*	*t Stat*	*P-value*
Intercept	65.09	10.22	6.37	0.0000
Avg. Pork Consumption	-0.19	0.16	-1.18	0.2563

a) Interprete los coeficientes de la regresión. ¿Le parecen intuitivos los signos?
b) ¿Son significativos? ¿En caso de serlo, a qué nivel de confianza?
c) Interprete el R2 de la regresión.

Imagine que ahora incorpora otra variable al modelo, el Precio de la Carne Vacuna, y corre nuevamente el modelo.

Avg. Annual Beef Consumption vs. Avg. Beef Price and Avg. Annual Pork Consumption

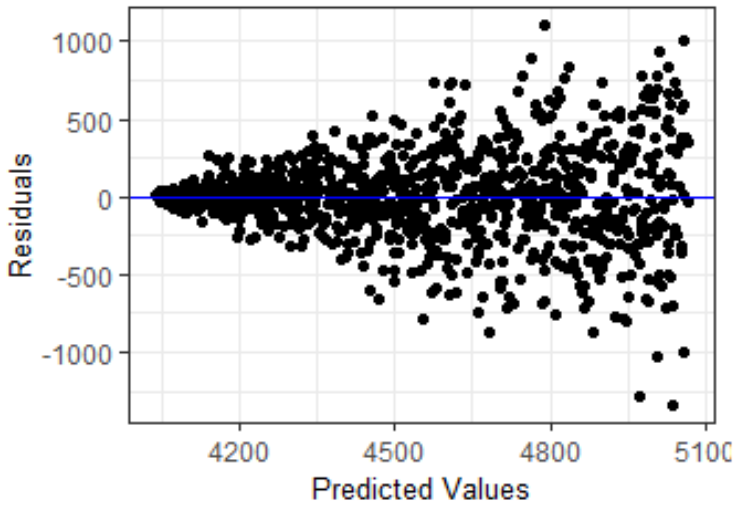
Regression Statistics	
Multiple R	0.9551
R^2	0.9123
Adjusted R^2	0.8998
Standard Error	1.3551
Observations	17

F test results	
F	*Signif.F*
72.81	0.0000

	Coefficients	*Std Error*	*t Stat*	*P-value*
Intercept	120.33	5.82	20.69	0.0000
Avg. Beef Price	-12.00	1.04	-11.49	0.0000
Avg. Pork Consumption	-0.40	0.05	-7.43	0.0000

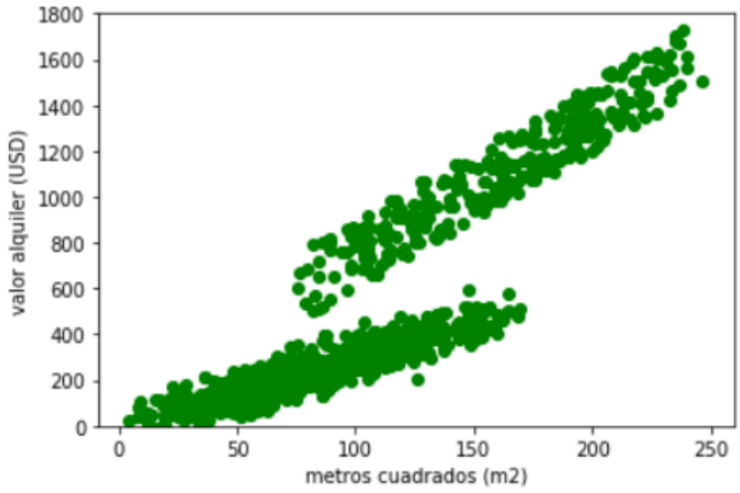
a) Interprete los coeficientes de la regresión. ¿Le parecen intuitivos los signos?
b) ¿Son significativos? ¿En caso de serlo, a qué nivel de confianza?
c) Interprete el R2 y R2 Ajustado de la regresión. ¿Cuál debería considerar y por qué?
d) ¿Por qué considera que cambió el coeficiente de la variable *Avg. Pork Consumption* de un modelo a otro?

11. Suponga que crea un modelo de regresión lineal y obtiene el siguiente gráfico de residuos:



a) En base al gráfico, ¿podríamos confiar en los resultados del modelo? Justifique su respuesta.
b) ¿Cómo deberían distribuirse los residuos en un modelo de regresión lineal?

12. Usted ha graficado un conjunto de datos que le han dado para entrenar modelos predictivos y para poder a futuro generar predicciones. Se sabe que los datos que ha graficado son muy representativos de la realidad. Usted únicamente conoce el modelo de regresión lineal. ¿Qué debe hacer en este caso? Argumente su respuesta.



13. Explique por qué al entrenar un modelo se divide el dataset en tres conjuntos: training, validation y testing, explicando qué es cada uno de estos conjuntos y qué problemas quiere evitar.