Walmart: Trip Type Classification

Ivonne, René, Ana e Ixchel

¿Cuántas personas compran diariamente en Walmart?

¡37 millones de clientes diarios!

Cifra más grande que la población de Canadá.

¿Qué sucedería si...?

Walmart encontrara una forma de capitalizar aún más su modelo de negocio a través de un entendimiento más profundo del comportamiento de los clientes. . .

kaggle

PLAN DE TRABAJO

no.	Actividades	Subtareas	Periodo	
1	Definir equipos	-	01-dic-18	01-dic-18
2	Definir base de datos a utilizar	Indagar en las reglas de las competencias de cada una de las posibles bases de datos Descargar las bases de datos Defnir el objetivo de cada base de datos	01-dic	07-dic
3	Estudiar la base de datos a detalle	Determinar los errores de registro para limpiar la base de datos Determinar si se cuenta con el poder de computo necesario para procesar los datos	07-dic	08-dic
		CRISP-DM		
4	Comprender el Negocio	Determinar los antecedentes de Walmart, es decir, el ambiente en el que se desarrolla y comprender su funcionamiento y cultura. Establecer el objetivo general y los objetivos específicos Determinar el criterio de éxito del proyecto (análisis de la competencia de kaggle para determinar medida de comparación) Establecer el plan del proyecto a detalle. Generar un documento en formato Rmarkdown con los elementos estudiados	08-dic	10-dic
5	Comprender los datos	Lectura de datos Evaluar con detalle los aspectos que deben limpiarse de la base de datos Generar un reporte reproduble que pueda ejecutar la lectura de datos y reporte los aspectos a limpia y considerar valores faltantes.	10-dic	11-dic
5	Preparar los datos	Seleccionar e integrar los datos Realizar la limpieza de datos (Quitar símbolos que entorpezcan el manejo de la base de datos, imputar datos, etc.) Realizar ingeniería de características Generar archivos reproducibles que faciliten la limpieza de datos y la ingeniería de características	11-dic	15-dic
6	Analizar los datos	Realizar el análisis univariado de los datos Realizar el análisis bivariado de los datos Realizar el análisis multivariado de los datos	15-dic	16-dic
7	Modelar en python	Proponer modelos a utilizar para cumplir con los objetivos del proyecto Modelar los datos utilizando python Correr modelos con datos de entrenamiento y prueba Comprar desempeño de modelos y ajustar	16-dic	19-dic
8	Evaluar el modelo	Seleccional el mejor modelo Comparar su desempeño en la competencia de Kaggle Mostrar lugar obtenido en la competencia	20-dic	20-dic
9	Generar un reporte final	Reportar todo lo enlistado con anterioridad Obtener conclusiones del proyecto	20-dic	20-dic
10	Preparar entregables	Elaborar una presentación que dure aproximadamente 15 minutos Acomodar archivos entregables Entregar todo lo necesario para cumplir con el objetivo del proyecto	20-dic	20-dic

1. Objetivos

Adquirir información relevante proveniente de las transacciones y otras variables para poder **clasificar a los clientes dependiendo de sus tipos de viaje** a las instalaciones de la empresa.

- ➔ **Análisis**
- ➔ **Modelado**
- ➔ **Selección de modelo**
- ➔ **Kaggle concurso**

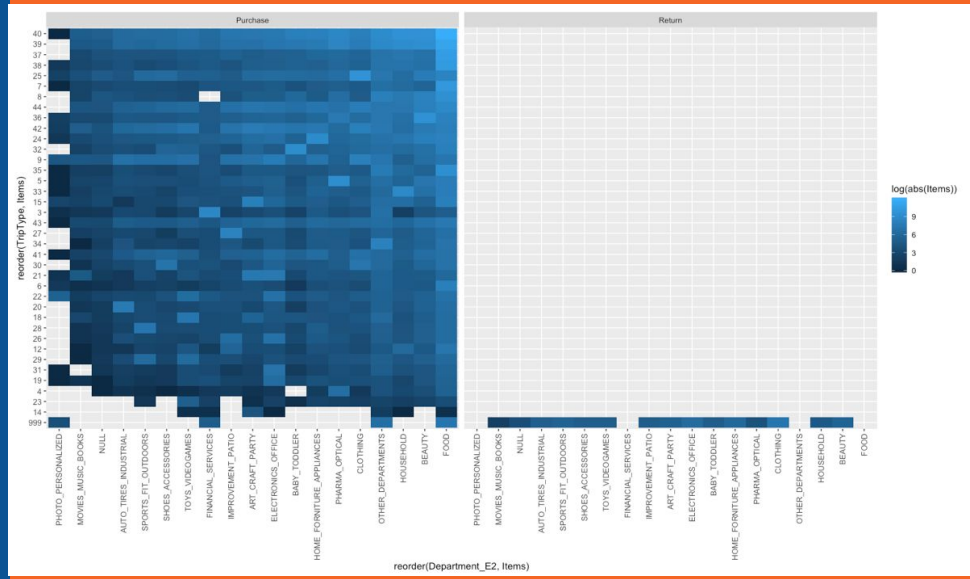
Metodología

Siguiendo **CRISP-DM** :

-Comprender que la prioridad es el cliente como individuo.

-Limpiar, analizar y crear datos que permitan identificar su motivo de visita.

-Desarrollar un algoritmo que predice la probabilidad individual del tipo de visita



Modelos

MODELO	ACCURACY / R CUADRADA	SCORE KAGGLE
GBOOST	0.6588	1.16162
REGRESIÓN LOGÍSTICA	0.6138	1.29374
EXTRA TREE CLASSIFIER	0.2623	-
RANDOM FOREST	0.42847	17.10050

Observaciones y recomendaciones

-Los datos provistos permiten explicar y pronosticar los patrones de consumo de productos

-Se requiere microdata del individuo para poder predecir mejor su motivo de visita (ej: hora de visita, tipo de pago, hora de salida, etc)

-Los resultados sugieren que la regresión logística y el GBoost resultan adecuados. Esto puede ser ya que captan la combinación de las preferencias por tipo de visita