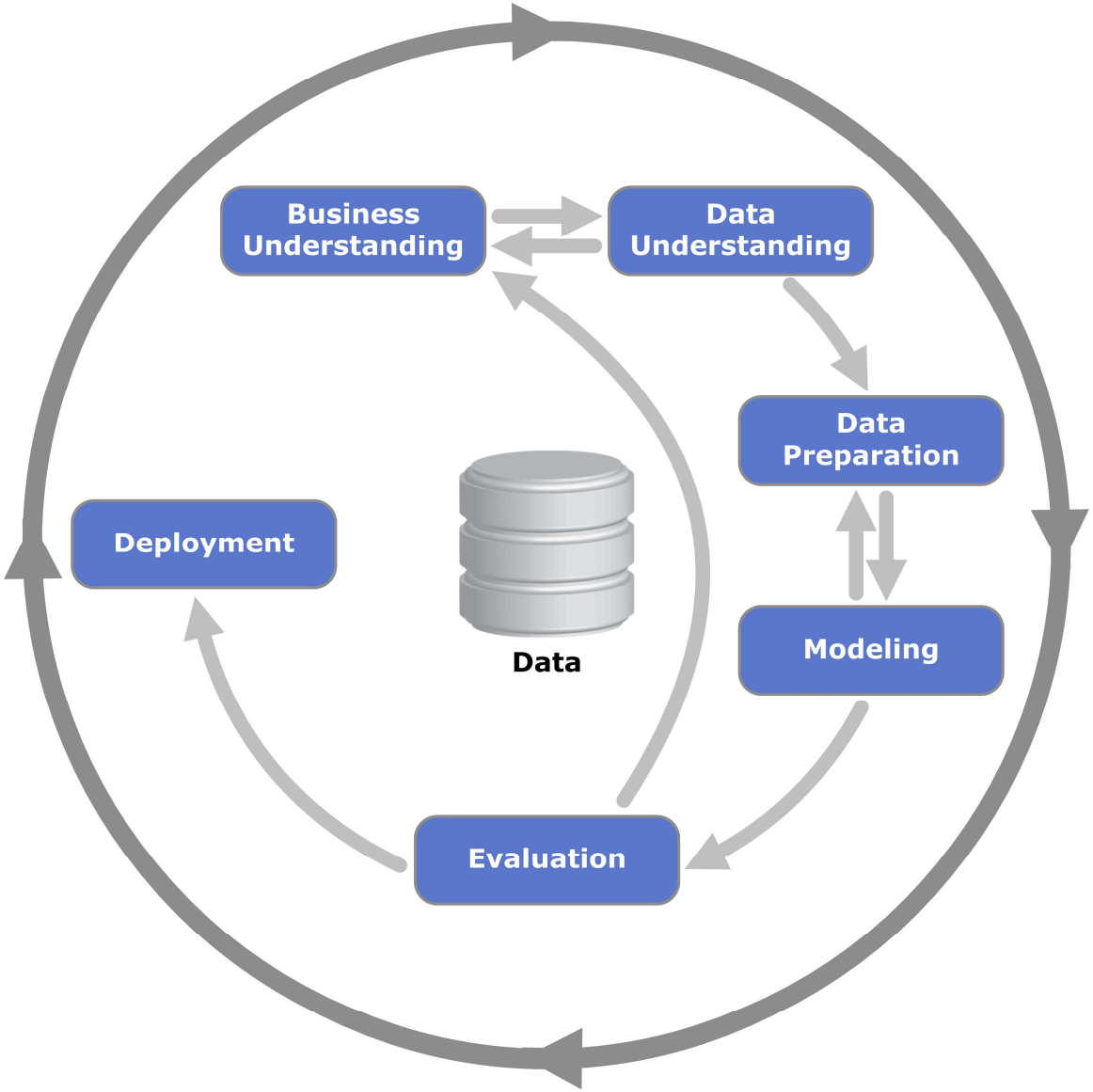
Clasificación de tipos de viajes Walmart

Metodología CRISP-DM

1. Comprensión del Negocio
2. Comprensión de los Datos
3. Preparación de los datos
4. Modelado
5. Evaluación
6. Implantación



Metodología CRISP-DM

1 Comprensión del Negocio

Objetivo

El objeto del presente trabajo es recrear la clasificación realizada por Walmart, mediante el empleo de técnicas de ciencia de datos, con el fin de ayudar a mejorar la toma de decisiones en cuanto al plan de mercadotecnia de la empresa y mejorar la experiencia de compra del cliente.

2 Comprensión de los datos

Calidad de los datos

Los datos aportados por Kaggle presentaron algunas inconsistencias como registros mal escritos, datos faltantes y no tenían formato Tidy.

EDA

Mediante el Análisis Exploratorio de Datos se pudo identificar a algunas de las variables relevantes que explican la clasificación de los viajes de los clientes en función del consumo.

3 Preparación de los Datos

Limpieza de los Datos

Mediante herramientas estadísticas se procedió a limpiar los datos capturados con el fin de prepararlos para la etapa de transformación.

Transformación de los Datos

Con el fin de obtener un mayor poder de predicción se le realizaron diversas transformaciones a los datos .

4 Modelado

Modelos

Con base en el desarrollo de distintos modelos estadísticos y de aprendizaje automático se buscó clasificar el tipo de viaje con base en las probabilidades obtenidas aplicando los modelos.

5 Evaluación

Función de Perdida

Evaluamos nuestros modelos con la función de perdida logarítmica multiclase.

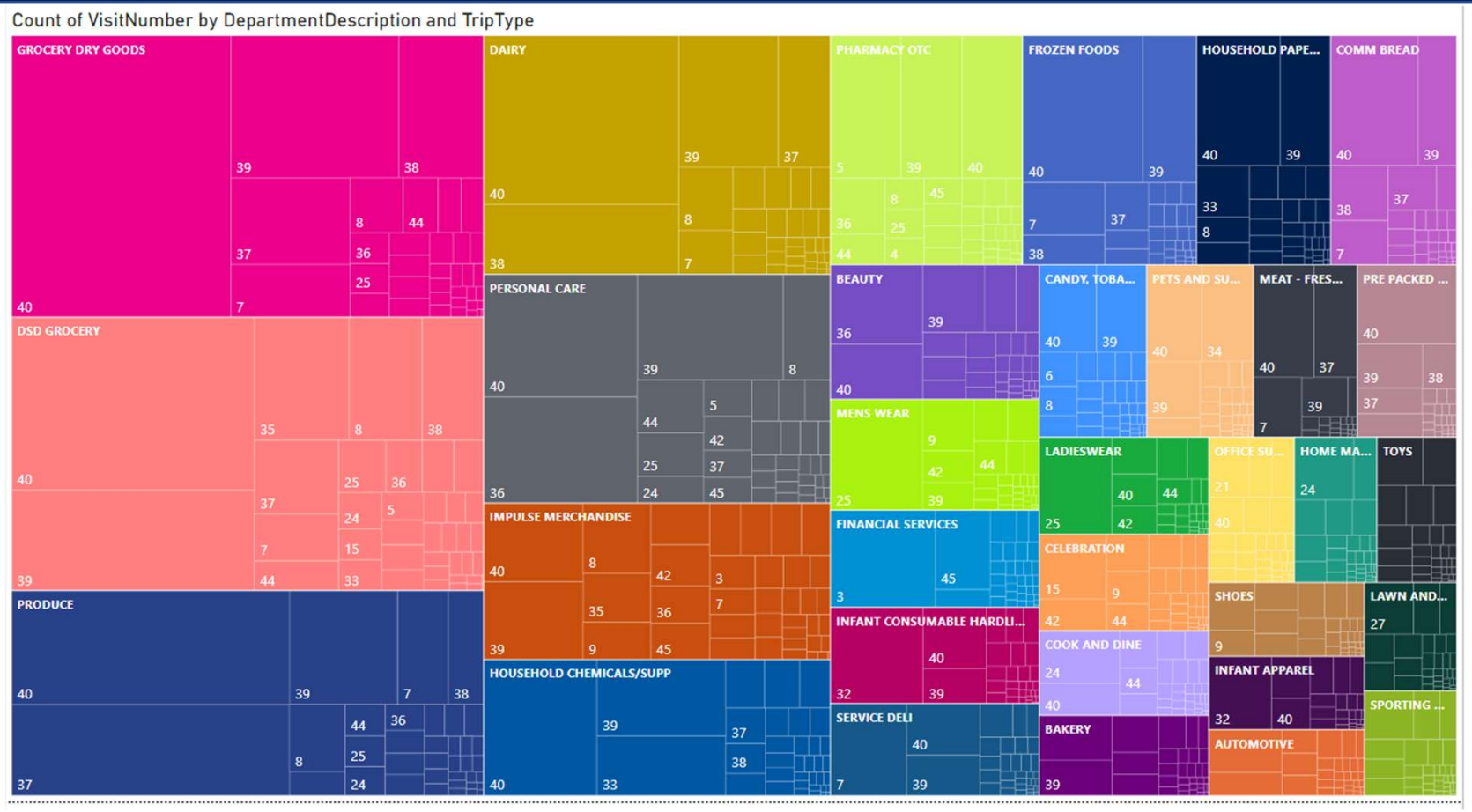
Comparamos los resultados de los modelos propuestos y elegimos el que tenía mejor puntuación.

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}),$$

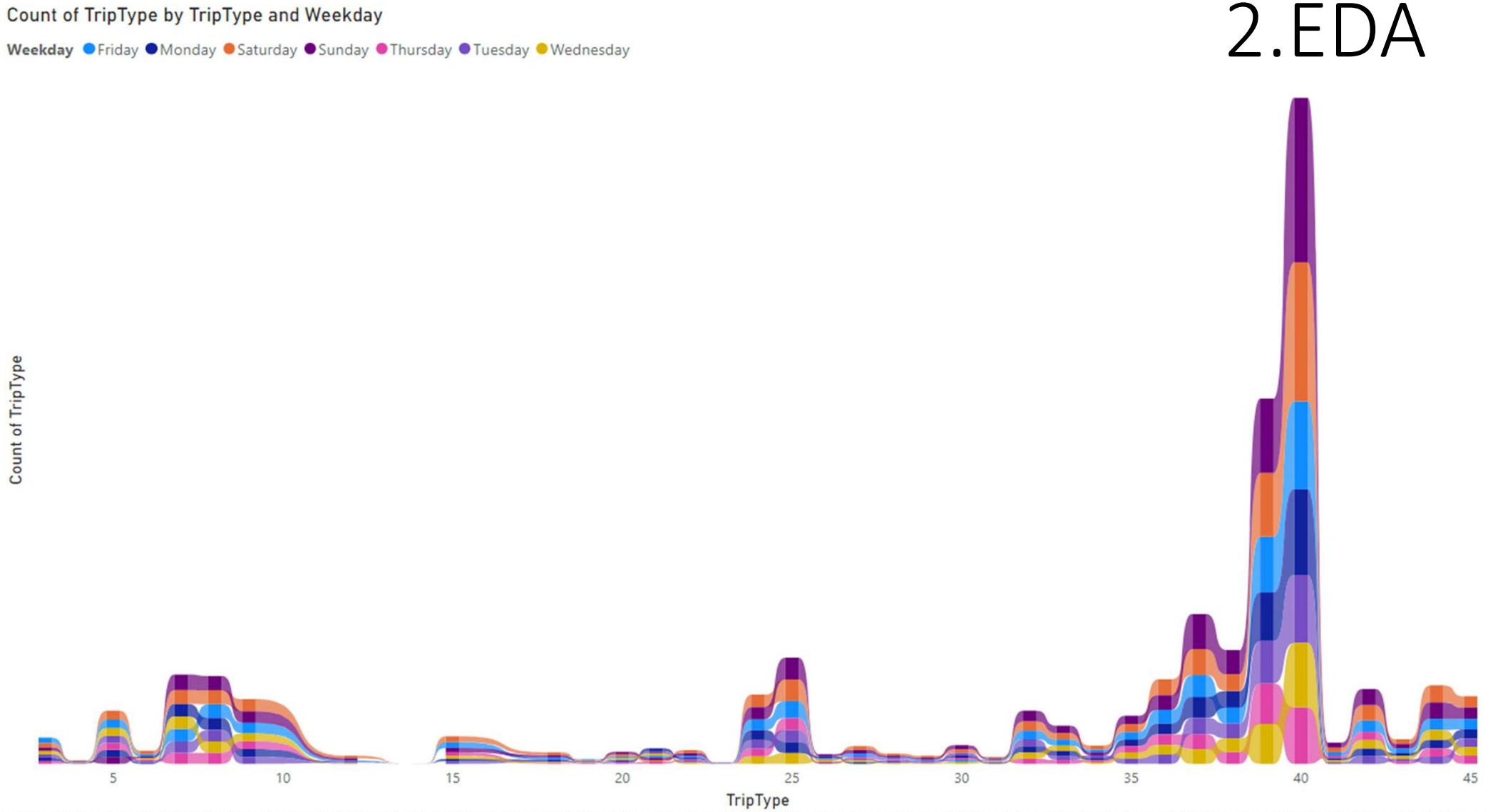
6. Implementación

Intentamos correr Flask pero por limitaciones de tiempo no se logró ese parte del proyecto.

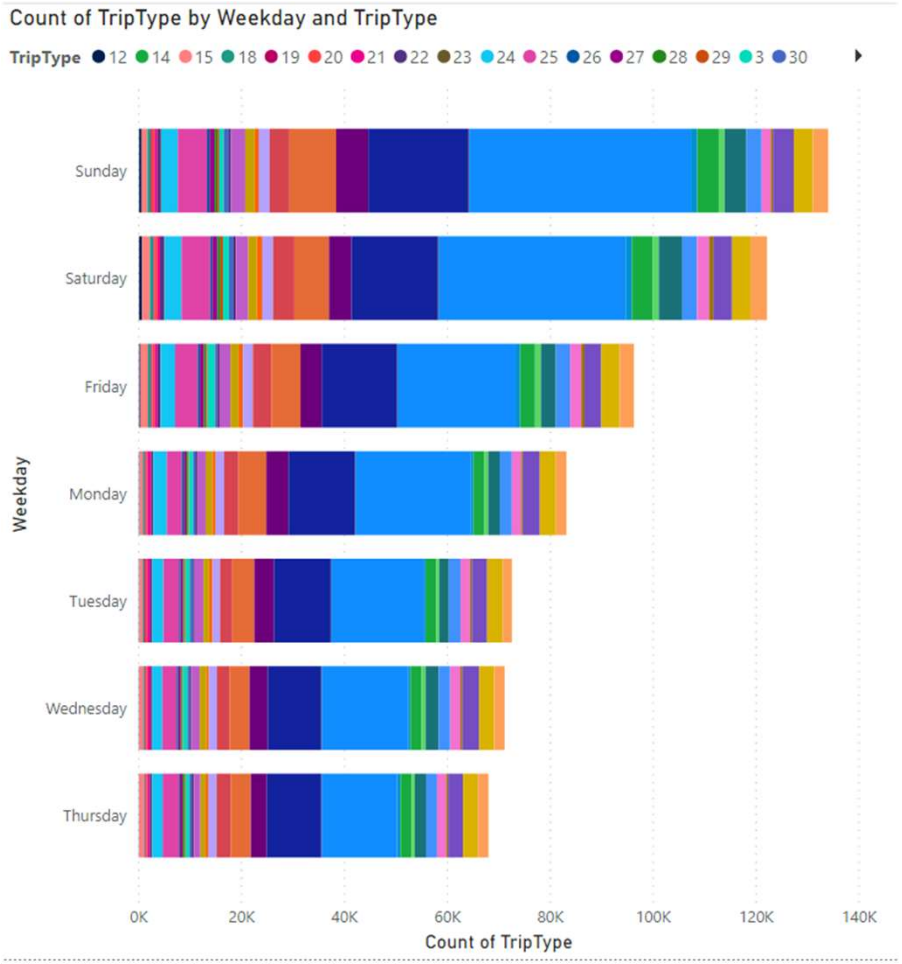
2 COMPRENSIÓN DE LOS DATOS

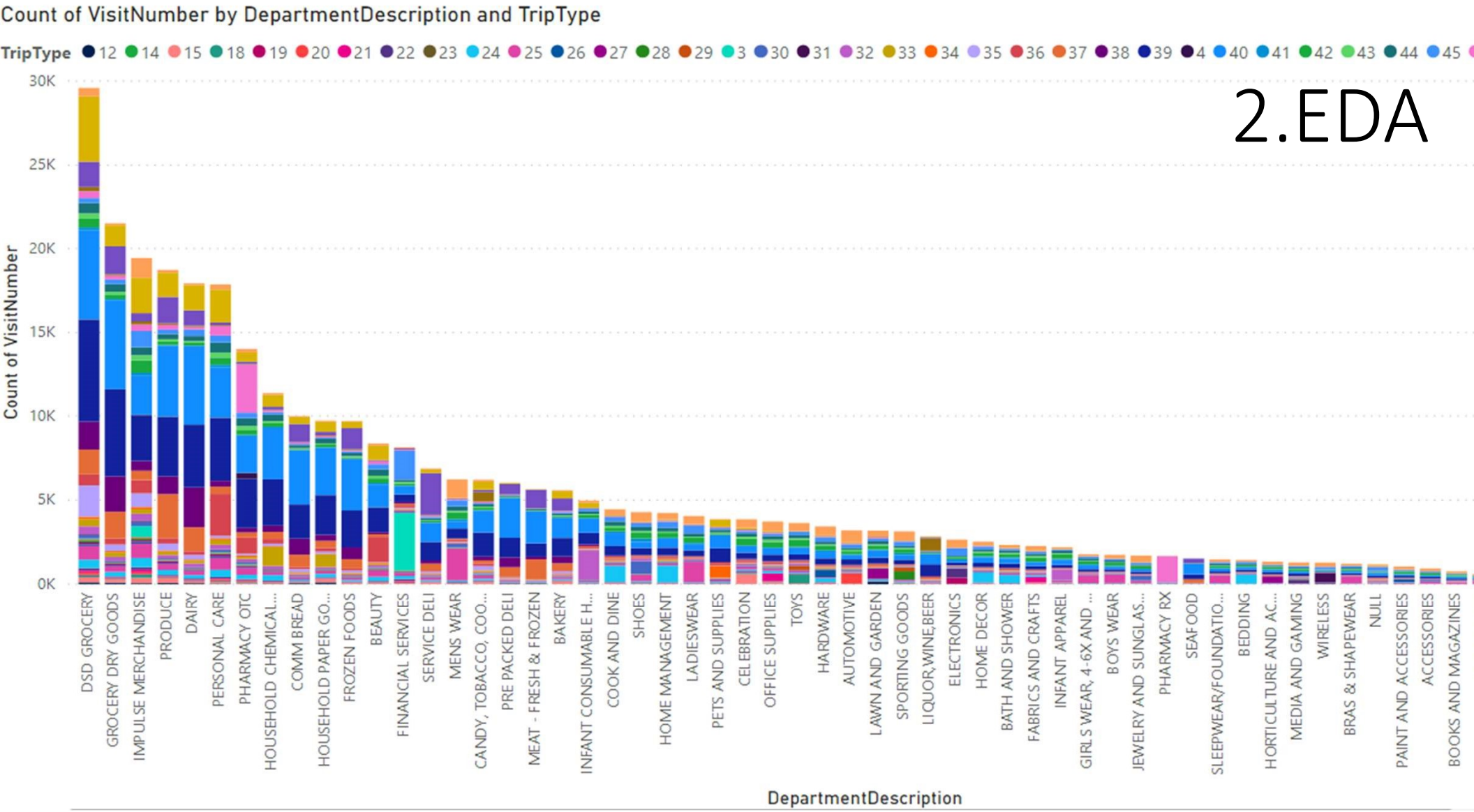


2.EDA

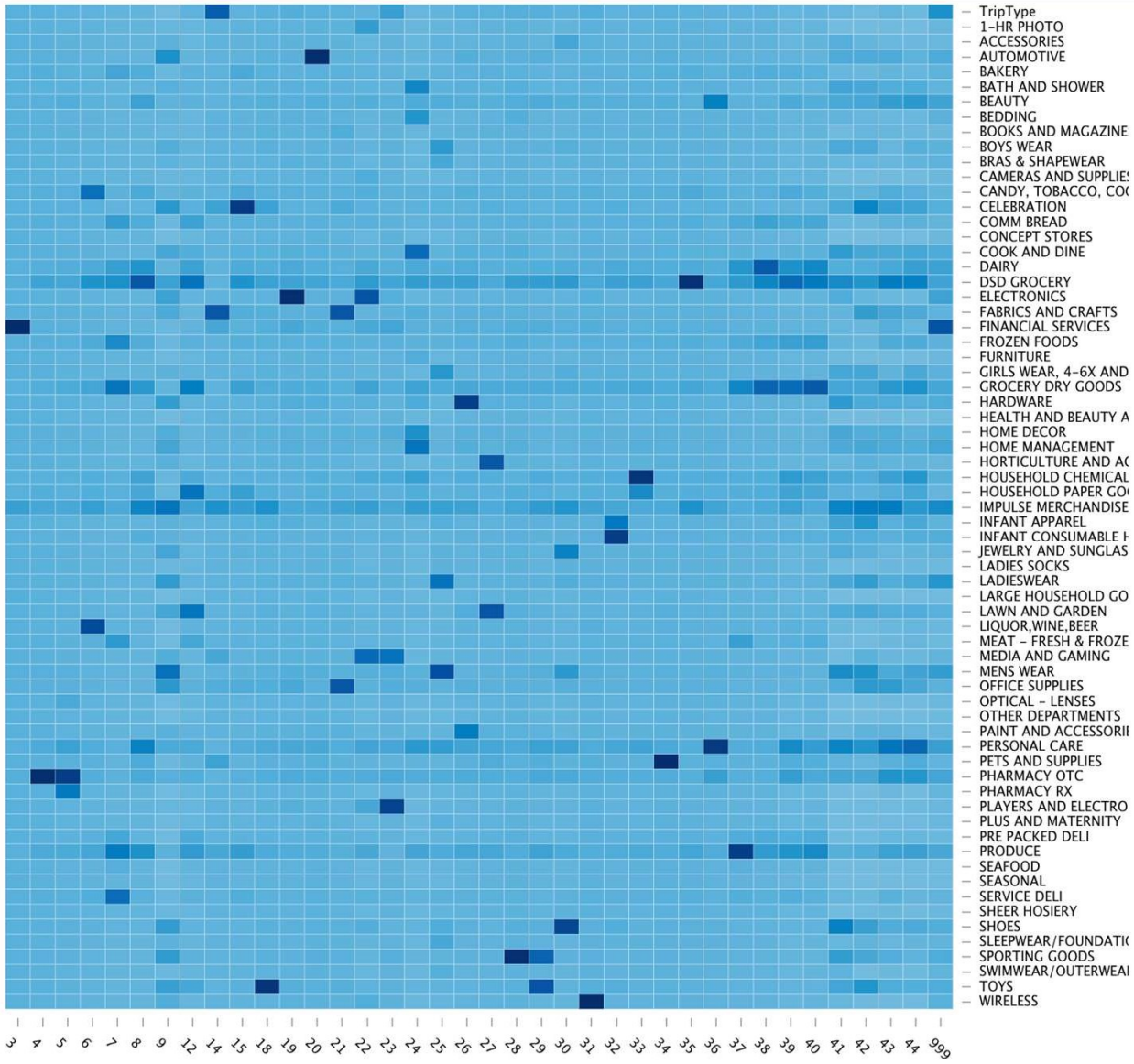

Count of TripType by TripType and Weekday
Weekday
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
Count of TripType
5
10
15
20
25
30
35
40
45
TripType

2.EDA


Count of TripType by Weekday and TripType
TripType 12 14 15 18 19 20 21 22 23 24 25 26 27 28 29 3 30
Sunday
Saturday
Friday
Monday
Tuesday
Wednesday
Thursday
Weekday
0K
20K
40K
60K
80K
100K
120K
140K
Count of TripType

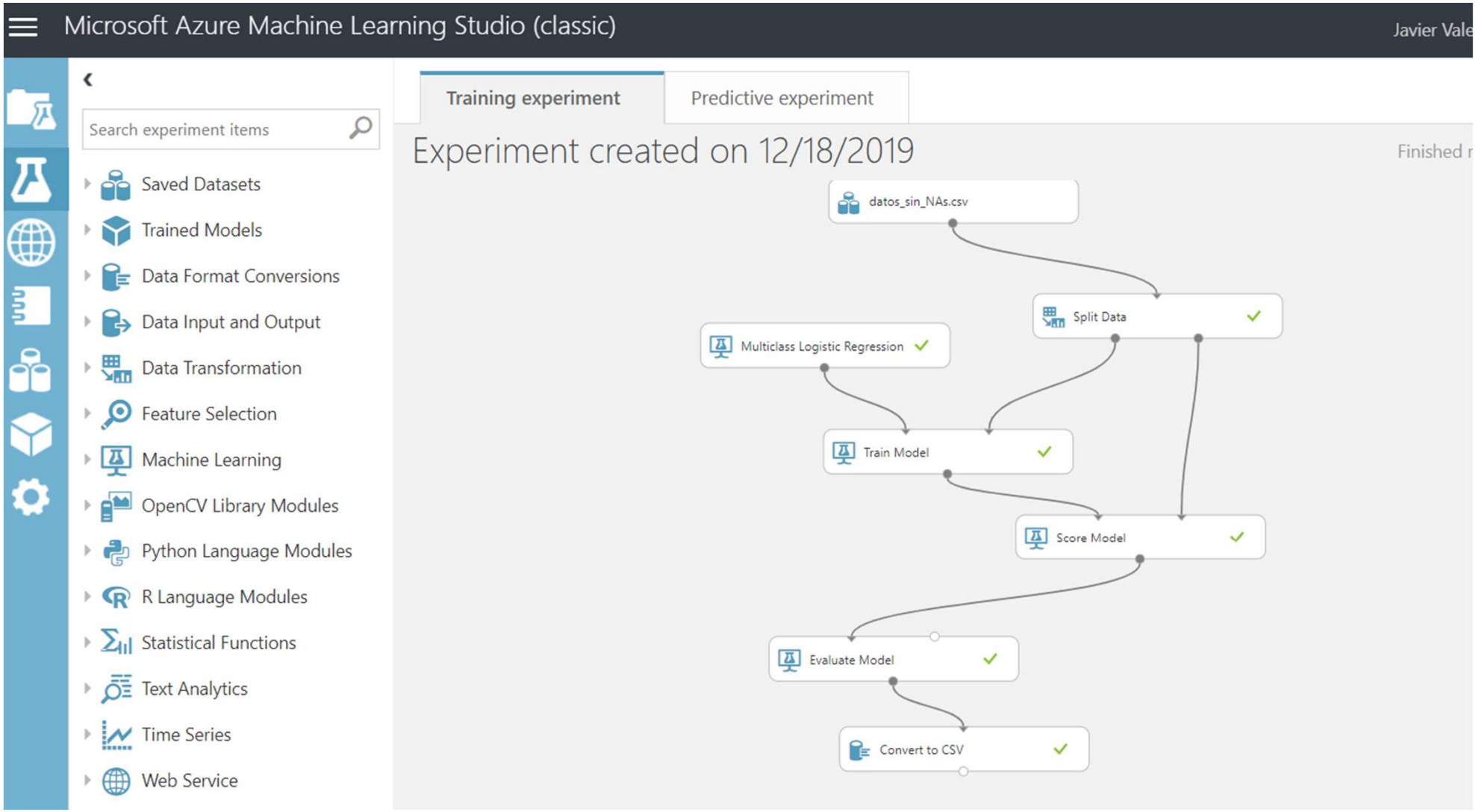

Count of VisitNumber by DepartmentDescription and TripType
TripType 12 14 15 18 19 20 21 22 23 24 25 26 27 28 29 3 30 31 32 33 34 35 36 37 38 39 4 40 41 42 43 44 45
Count of VisitNumber
30K
25K
20K
15K
10K
5K
0K
DSD GROCERY
GROCERY DRY GOODS
IMPULSE MERCHANDISE
PRODUCE
DAIRY
PERSONAL CARE
PHARMACY OTC
HOUSEHOLD CHEMICAL...
COMM BREAD
HOUSEHOLD PAPER GO...
FROZEN FOODS
BEAUTY
FINANCIAL SERVICES
SERVICE DELI
MENS WEAR
CANDY, TOBACCO, COO...
PRE PACKED DELI
MEAT - FRESH & FROZEN
BAKERY
INFANT CONSUMABLE H...
COOK AND DINE
SHOES
HOME MANAGEMENT
LADIESWEAR
PETS AND SUPPLIES
CELEBRATION
OFFICE SUPPLIES
TOYS
HARDWARE
AUTOMOTIVE
LAWN AND GARDEN
SPORTING GOODS
LIQUOR,WINE,BEER
ELECTRONICS
HOME DECOR
BATH AND SHOWER
FABRICS AND CRAFTS
INFANT APPAREL
GIRLS WEAR, 4-6X AND ...
BOYS WEAR
JEWELRY AND SUNGLAS...
PHARMACY RX
SEAFOOD
SLEEPWEAR/FOUNDATIO...
BEDDING
HORTICULTURE AND AC...
MEDIA AND GAMING
WIRELESS
BRAS & SHAPEWEAR
NULL
PAINT AND ACCESSORIES
ACCESSORIES
BOOKS AND MAGAZINES
DepartmentDescription

2.EDA



3. PREPARACIÓN DE LOS DATOS

4. MODELADO


Microsoft Azure Machine Learning Studio (classic)
Javier Vale
Search experiment items
Saved Datasets
Trained Models
Data Format Conversions
Data Input and Output
Data Transformation
Feature Selection
Machine Learning
OpenCV Library Modules
Python Language Modules
R Language Modules
Statistical Functions
Text Analytics
Time Series
Web Service
Training experiment
Predictive experiment
Experiment created on 12/18/2019
Finished r
datos_sin_NAs.csv
Split Data
Multiclass Logistic Regression
Train Model
Score Model
Evaluate Model
Convert to CSV

5. EVALUACIÓN Y RESULTADOS

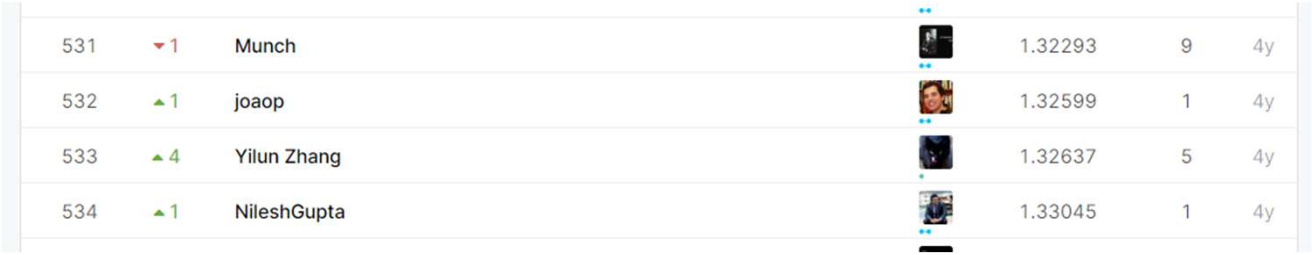
9 submissions for **Daniela Pinto Veizaga** Sort by Private Score

All **Successful** Selected

Submission and Description	Private Score	Public Score	Use for Final Score
pre_GB_std.csv 6 hours ago by Daniela Pinto Veizaga add submission details	1.32845	1.33002	☐
pre_GB.csv 6 hours ago by Daniela Pinto Veizaga XGboost	1.40185	1.40690	☐
pre_logit.csv 8 hours ago by Daniela Pinto Veizaga third try	1.42031	1.42349	☐

6. CONCLUSIONES

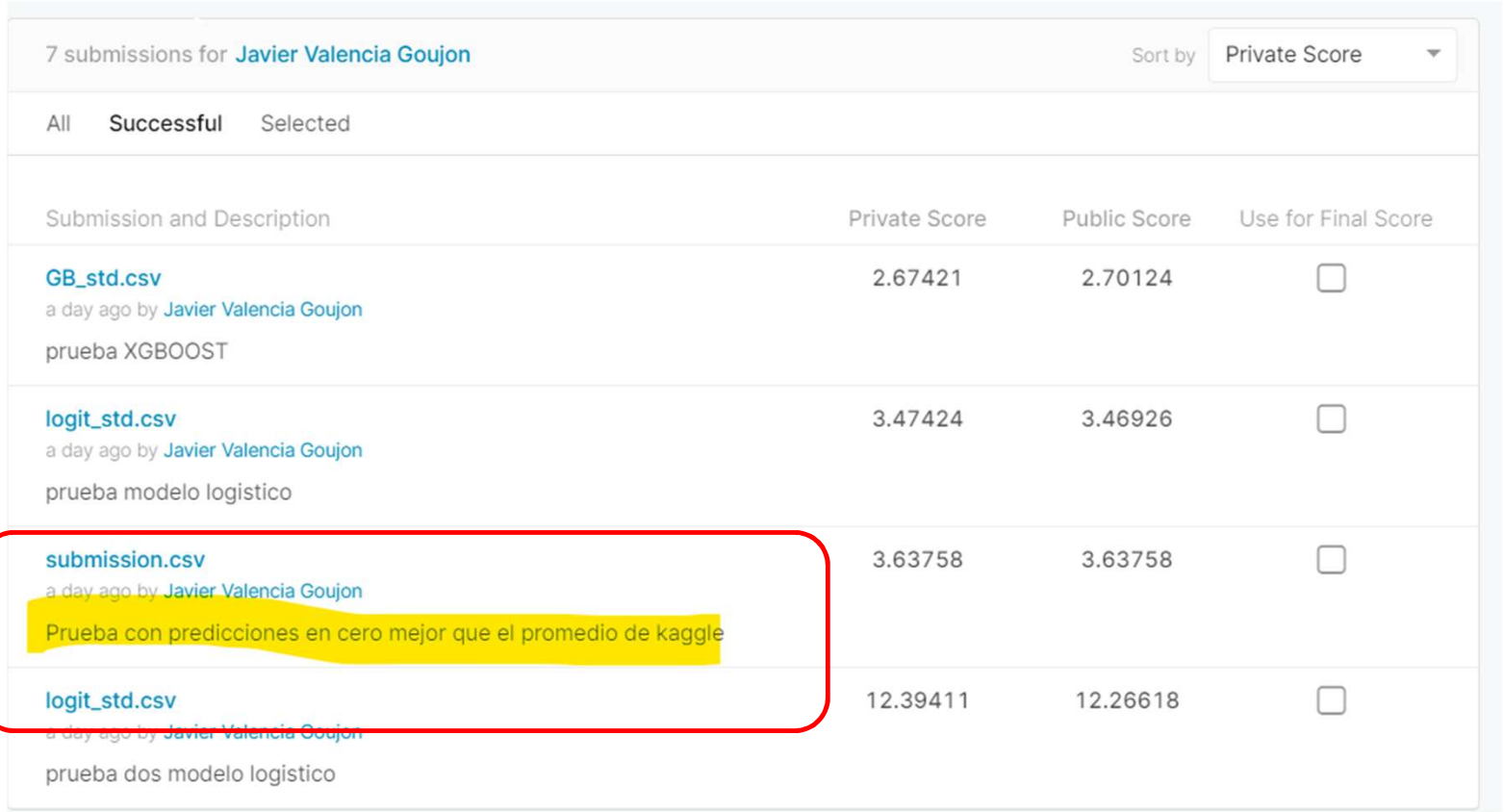
1. Feature Engineering fue clave para resolver este problema, pues permitió reducir el error del modelo significativamente.
2. El problema más grande encontrado fue el tamaño del dataset, lo cual nos dificulto un poco hacer las transformaciones necesarias.
3. Usando herramientas como Azure ML Studio facilita y reduce el tiempo para entrenar y probar los modelos.
4. El mejor modelo fue gradient boosting después de aplicar feature engineering con un error de 1.32 lo cual nos dejo en el lugar 534



531	▼1	Munch	1.32293	9	4y
532	▲1	joaop	1.32599	1	4y
533	▲4	Yilun Zhang	1.32637	5	4y
534	▲1	NileshGupta	1.33045	1	4y

BONUS SLIDES - ANEXO

Hackeando Kaggle: resultados interesantes

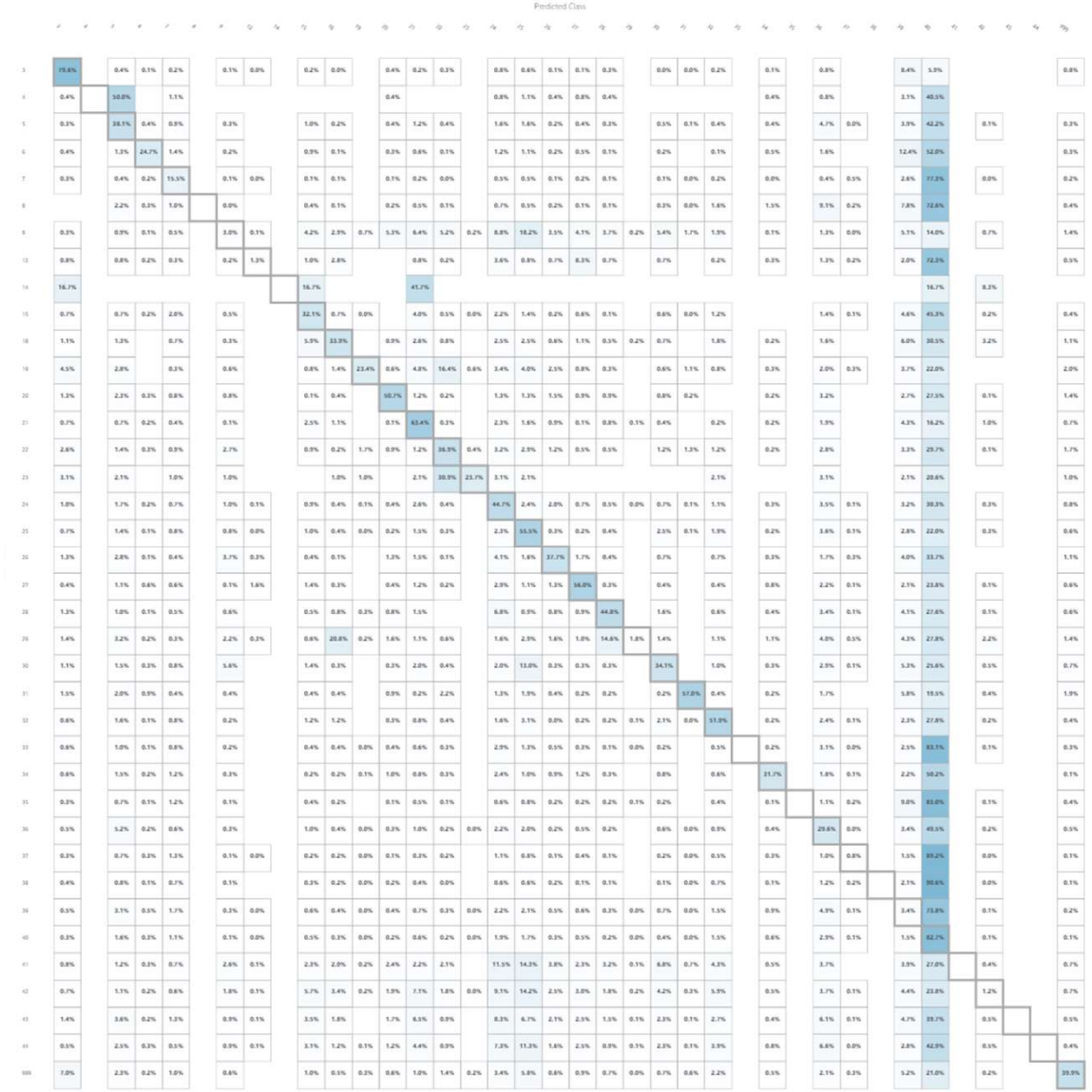


7 submissions for Javier Valencia Goujon

Sort by Private Score

All Successful Selected

Submission and Description	Private Score	Public Score	Use for Final Score
GB_std.csv a day ago by Javier Valencia Goujon prueba XGBOOST	2.67421	2.70124	☐
logit_std.csv a day ago by Javier Valencia Goujon prueba modelo logistico	3.47424	3.46926	☐
submission.csv a day ago by Javier Valencia Goujon Prueba con predicciones en cero mejor que el promedio de kaggle	3.63758	3.63758	☐
logit_std.csv a day ago by Javier Valencia Goujon prueba dos modelo logistico	12.39411	12.26618	☐



Resultados
Azure ML Studio

Metrics

Overall accuracy	0.348963
Average accuracy	0.965735
Micro-averaged precision	0.348963
Macro-averaged precision	NaN
Micro-averaged recall	0.348963
Macro-averaged recall	0.262886