# Resumen

Este documento habla del proceso de entrenar una modelo de regresión para estimar los tiempos de despacho de las ordenes realizadas a Cornershop.

# Generación del Set de Datos

La generación de los datos consistió en correlacionar los datos da cada archivo con su orden correspondiente y obtener los datos relevantes para realizar regresión.

|  |  |  |
| --- | --- | --- |
| Archivo | dato extraido | comentarios |
| Orders.csv | lat y lng  dow  on\_demand  total\_minutes | Se utilizó la posición del usuario final, el día que la orden fue realizada (dow), si es que el pedido fue solicitado en menos de 90 minutos (on\_demand) y el tiempo que demoró la entrega de cada orden (total\_minutes). |
| order\_product.csv | num\_products\_KG  num\_products\_UN | Para cada orden, se obtuvo el número de productos comprados que se pesan y los que son vendidos por unidad. |
| shoppers.csv | picking\_speed | Velocidad de compra asociada al shopper que realizó la orden |
| storebranch.csv | lat y lng | Dirección del supermercado |

# Transfromaciones y Preprocesamiento de Datos

Para obtener utilidad de los datos en formato latitud y longitud, se utilizó la API de Google Maps (https://maps.googleapis.com/maps/api/distancematrix/<request>) para obtener el tiempo estimado por Google entre dos lugares, utilizando un auto y sin considerar el tráfico. Para obtener los datos se creó el script googlemapsutil.py. De Google también se podría obtener la distancia o el tiempo estimado considerando el tráfico si se desea.

Finalmente los siguientes datos fueron utilizados para entrenar el modelo:

|  |  |  |
| --- | --- | --- |
| Archivo | Columnas | Comentarios |
| pd\_data\_train.csv | google\_eta  num\_products\_KG  num\_products\_UN  picking\_speed  dow  on\_demand  total\_minutes | total\_minutes tiene un valor en todas las filas de este archivo. |
| pd\_data\_unknown.csv | google\_eta  num\_products\_KG  num\_products\_UN  picking\_speed  dow  on\_demand  total\_minutes | total\_minutes está vacío en todas las filas de este archivo. |

# Entorno y Herramientas Utilizadas

Se utilizaron las siguientes librerías de python.

|  |  |
| --- | --- |
| Librería | Comentarios |
| pandas | Para trabajar con DataFrames. Transformaciones y filtros de datos |
| numPy | Necesario para usar pandas y sklearn |
| sklearn | Para realizar regresión y evaluar el modelo |
| pickle | Para guardar el modelo entrenado |
| matplotlib | Para estudiar gráficos de los resultados del modelo de regresión |

# Modelos de Predicción Utilizados

|  |  |
| --- | --- |
| Nombre | Comentarios |
| **sklearn.linear\_model. LinearRegression** | **(modelo elegido)** Dada la naturaleza del problema, regresión lineal es un modelo simple y efectivo. |
| sklearn.linear\_model.Lasso | Similares resultados a LinearRegression |
| sklearn.svm.LinearSVR | Similares resultados a LinearRegression |
| sklearn.svm.SVR | Similares resultados a LinearRegression |
| sklearn.NuSVR | Similares resultados a LinearRegression |

Los modelos distintos a Regresión Lineal fueron probados durante el proceso, para probar si podrían otorgar mejores resultados. Como no fue el caso, se dejaron a un lado.

# Evaluación del modelo

Se entrenó la regresión lineal con un 70% y 90% de los datos; el resto era utilizado para probar la calidad del modelo.

Se utilizó para adquirir una idea de cuán bueno es cada modelo. También se observan ‘a mano’ los datos predichos vs los reales para obtener una idea de las estimaciones y los errores.

El error promedio de las estimaciones fue grande, de 22 minutos. Este valor es obtenido utilizando sklearn.metrics. mean\_squared\_error, de sklearn.

En el archivo src/classifiers/LinearRegression000.txt se guardaron los resultados e información de la regresión lineal.

# Predicciónes de valores

En la carpeta src/predictions/ hay dos archivos con predicciones realizadas por el modelo entrenado. Las predicciones son exclusivamente las que estaban faltando en el archivo original *orders.csv.*

|  |  |
| --- | --- |
| Nombre Archivo | Comentarios |
| predictions.csv | Igual al archivo /data/pd\_data\_unknown.csv pero con la última columna NO vacía |
| orders\_with\_predictions.csv | Copia del archivo /data/orders.csv pero la última columna solo tiene valores en las filas donde en el original no las tenía. |