Curso Data Science

Proyecto final | Modelo de Predicción de Órdenes

Comisión 25570

2022

Entrega final

Integrantes:

Federico Campana

Fernando Martínez

Tutor:

Mathiews Zavala

Profesor:

Damián Dapueto

Índice

SISTEMA PREDICTIVO DE ÓRDENES

Introducción

Uno de los grandes requerimientos de las empresas hoy en día es tratar de predecir ciertos aspectos de sus negocios para poder hacer más efectivos los recursos de las mismas empresas.

Un ejemplo de esta necesidad es saber cuán requerido será el servicio que ofrecen diariamente las empresas de delivery, por lo que nos generó la duda de si se puede desarrollar un sistema de predicción de entregas y de cantidad de cadetes necesarios para cubrir esas entregas. Es por esto que se tomó la empresa Pedidos Ya como caso de ejemplo para poder realizar un desarrollo de este tipo.

Pedidos Ya nace en el año 2008 como parte de un proyecto universitario en el cual un grupo de tres amigos debían presentar una idea de negocios en 15 minutos. La idea de negocios era crear una web en la cual se pueda pedir "chivitos", el famoso sándwich uruguayo. En el año 2012 desembarca en Argentina con un único competidor, Rappi. En el año 2014 se une con la plataforma líder mundial en delivery online, Delivery Hero, lo que permitió impulsar el negocio. Por último, en el año 2018 implementa su propio servicio de logística para resolver la creciente demanda en los diferentes mercados.

Con la implementación del servicio de logística propio de Pedidos Ya se creó una nueva necesidad: poder calcular efectivamente la cantidad de órdenes que se pueden generar en un día, dividido por zona geográfica, y con esto poder hacer un cálculo de la cantidad de cadetes necesarios para cubrir esa demanda. Pero en esta necesidad hay una variable que no es tan predecible, el clima. Esta última variable es, en cierta medida, impredecible lo que hace que este cálculo no sea exacto y genera que haya gastos extras por parte de la empresa que pueden ser evitados.

Temática

Este razonamiento nos genera la siguiente pregunta: con los datos recopilados de órdenes, métricas logísticas y un historial del clima, ¿se puede desarrollar un sistema que cruce la información del pronóstico del clima con el historial de órdenes y nos dé un estimado de órdenes para los días siguientes que sea casi exacto?

Objetivo

Nuestro objetivo es poder determinar en base a un análisis de los datos si se pueden predecir las órdenes en base a los datos históricos, teniendo en cuenta variables como el clima y los eventos importantes.

Nuestro Equipo

Federico Campana (líder): Tengo 30 años, soy de Buenos Aires, Argentina. Soy licenciado en administración de empresas y me desempeño como Controller financiero para ExxonMobil.

Fernando Martinez: Tengo 29 años, soy de Córdoba Capital, Argentina y si bien soy licenciado en turismo me desempeño como Business Analyst para el equipo Regional de Pedidos Ya.

Fuente de Trabajo

Nuestro dataset contiene información de la ciudad de Córdoba de todo el año 2021 y hasta agosto de 2022.

La información con la que contamos consta de:

- Fecha: división de los datos por día
- Count_riders: la cantidad de cadetes disponibles por día, zona y vehículo
- Confirmed_orders: órdenes confirmadas
- Cancelled_orders: órdenes canceladas
- AVG_delivery_time: tiempo de entrega promedio
- AVG_dropoff_distance: distancia promedio entre el restaurante y el usuario
- AVG_acceptance_rate: tasa de aceptación de órdenes por los cadetes
- Working_hours: cantidad de horas trabajadas por los cadetes
- Shifts_done: cantidad de turnos trabajados (cada turno puede ser un mínimo de 2 horas con máximo de 6 horas)
- Event_type: señala si en el día hubo algún evento deportivo importante
- Precipitaciones en mm: nivel de precipitación del día
- Temp Max: temperatura máxima registrada
- Temp min: temperatura mínima registrada
- Presión Atm: Presión atmosférica
- %de Hr: % de horas del día con cielo cubierto por nubes
- Nubosidad: nivel de nubosidad, siendo 1 muy bajo y 8 cielo totalmente cubierto.

Se utilizaron estas variables debido a que, con los datos de órdenes, riders y métricas logísticas se puede hacer una previsión de la cantidad de riders que se necesitan para cubrir la cantidad de órdenes. Por otro lado, la selección de las variables climáticas parte de ser las variables necesarias para que haya lluvia.

Objetivo General

Nuestro objetivo para estos datos es poder, por un lado, analizar la relación entre las zonas, las cantidades de órdenes y métricas logísticas y, por otro lado, analizar la relación entre las variables climáticas y las precipitaciones.

Cruzando estas variables es cómo determinamos que se podría obtener la predicción de órdenes y luego de esto, poder determinar la cantidad de oferta de cadetes que debería haber para cada zona logística.

Objetivos Específicos

Como objetivos específicos tenemos:

- Desarrollar las instancias de Data Acquisition y Data Wrangling
- Definir nuestra variable target
- Seleccionar nuestros algoritmos candidatos, definiendo parámetros y probar, entre los distintos candidatos, cuál es el algoritmo que más se ajusta a nuestra necesidad.

Como primer objetivo, realizamos un análisis descriptivo de nuestros datos para analizar el tipo de dato que tiene nuestro dataset e identificar si hay que realizar alguna conversión del tipo de dato, analizamos la cantidad de valores nulos para trabajarlos y realizamos un análisis descriptivo de las variables.

Una vez realizado estos pasos básicos procedemos a crear dos variables que nos ayudarán a nuestro modelo predictivo. La primera variable es la cantidad de órdenes por rider (realizando la cantidad de órdenes confirmadas dividido la cantidad de riders disponibles por día) y la segunda, la amplitud térmica, la cual se obtiene de la resta entre la temperatura máxima y la mínima.

Estas dos variables nos van a permitir por un lado analizar la cantidad de órdenes promedio que puede entregar cada cadete y las diferencias entre las temperaturas respectivamente. Los cambios según la estación de estas dos variables nos pueden hacer entender un poco mejor si una es correlativa de la otra.

Como segundo objetivo, nosotros lo que intentamos predecir es la cantidad de órdenes que vamos a tener un día específico, por lo que determinamos que nuestra variable target va a ser la de “confirmed_orders”.

Ya que definimos nuestra variable realizamos distintos análisis entre esta variable y las demás para determinar la correlación de las mismas o si hay tendencias entre las distintas variables. Para lograr esto se realizaron distintos gráficos para su análisis:

- Gráfico lineal entre la cantidad de órdenes confirmadas y el mes del año: con este gráfico podemos determinar si hay alguna tendencia de compra durante el transcurso del año.
- Histograma entre las temperaturas máximas y mínimas para ver cuáles son las temperaturas más frecuentes
- Heatmap y gráficos de correlación de pearson, kendall y spearman para poder analizar las correlaciones entre las distintas variables y determinar si algunas variables tienen correlación o no y si esta correlación es positiva o negativa.
- pointplot conjunto entre la cantidad de órdenes confirmadas y la cantidad de cadetes para ver cómo es su relación

Luego de estos gráficos básicos procedemos a realizar un forecasting de la cantidad de órdenes con la librería prophet. Se eligió este algoritmo ya que al tener un dataset que utiliza series de tiempo este algoritmo mencionado es el ideal para procesar los datos.

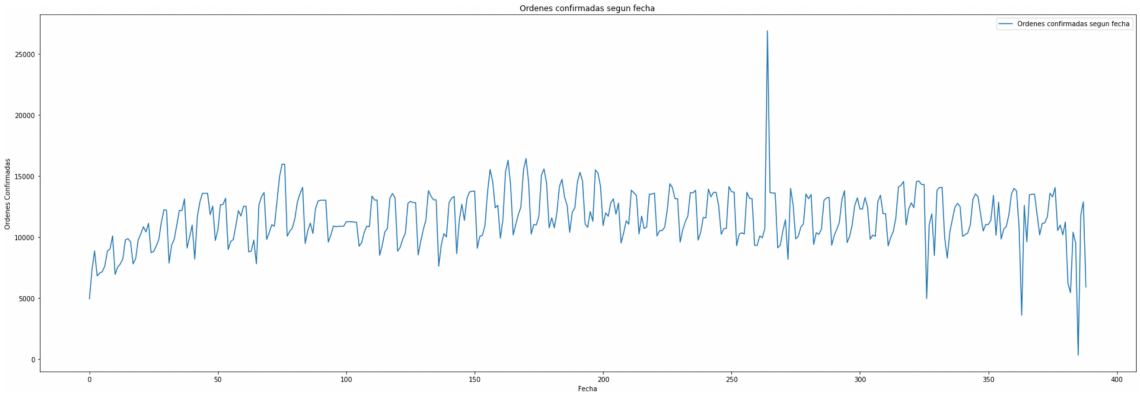
Con este algoritmo no solo realizamos un pronóstico de la cantidad de órdenes, sino también de la cantidad de riders que vamos a necesitar para cubrir estas órdenes y no solo se realizó un pronóstico a nivel mensual, sino también se realiza un pronóstico a nivel semanal para poder determinar las tendencias de estas dos variables durante los días de la semana.

Por otro lado, se realizó un Decision Tree Regresor en la cual no se tiene en cuenta la serie de tiempo, pero se puede realizar una predicción de las órdenes en base a un entrenamiento de la historia de las órdenes, día por día.

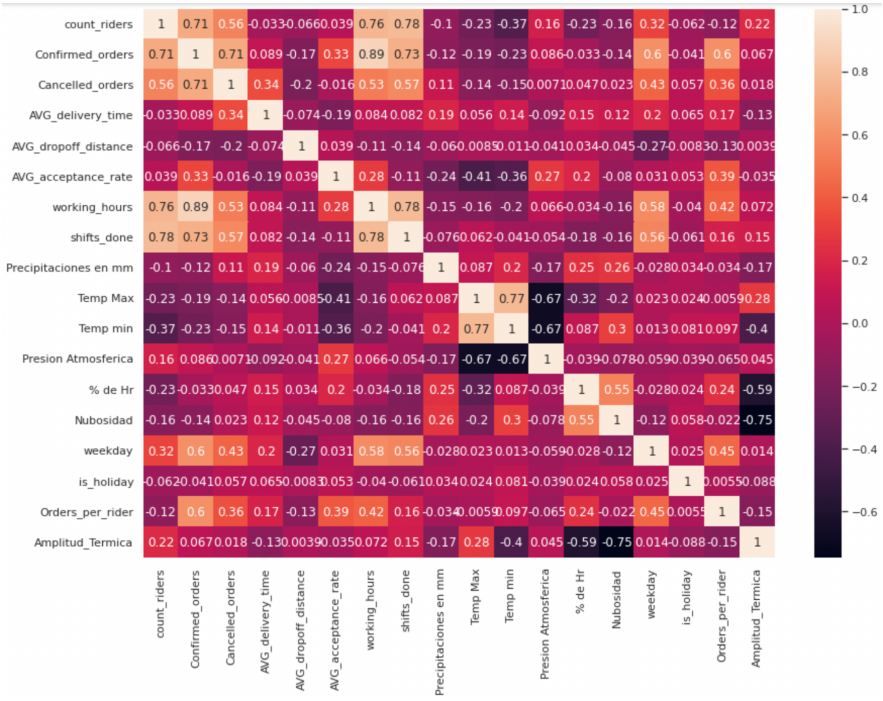
Luego de realizar el modelo del árbol de decisión, el puntaje de este algoritmo da un valor de 0.99.

Análisis exploratorio de datos

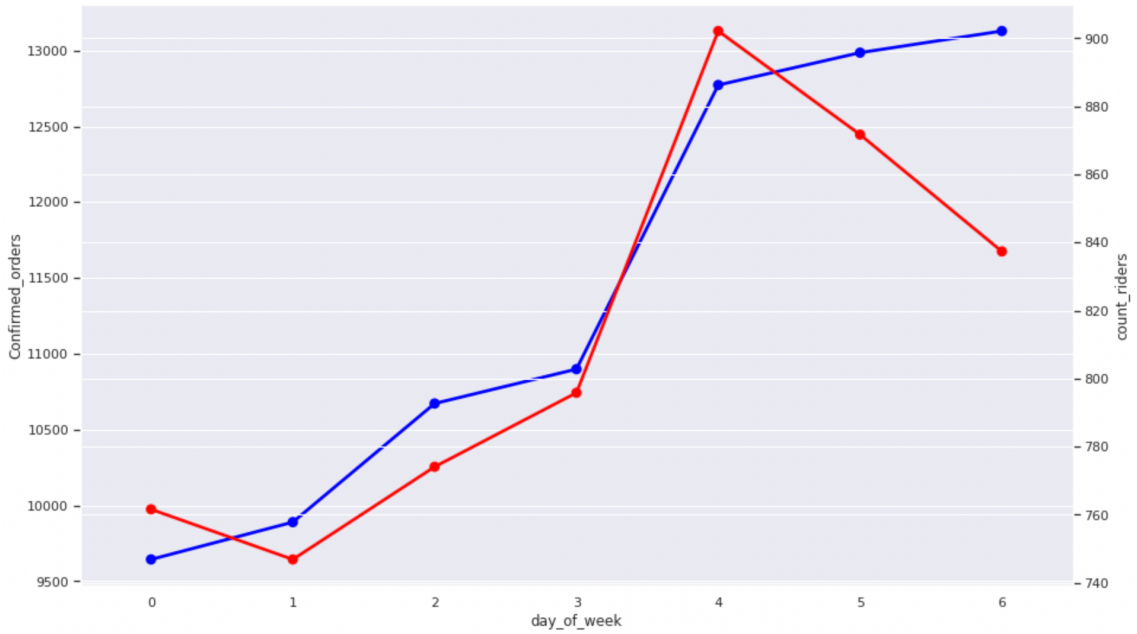
Con el siguiente gráfico de línea se puede observar la tendencia de los pedidos recibidos en la ciudad de Córdoba.



Mapa de calor de correlación de todas las variables



Se puede observar cómo se distribuyen los pedidos según el día de la semana, siendo la mayor parte entre los días viernes y domingos.



Algoritmos elegidos

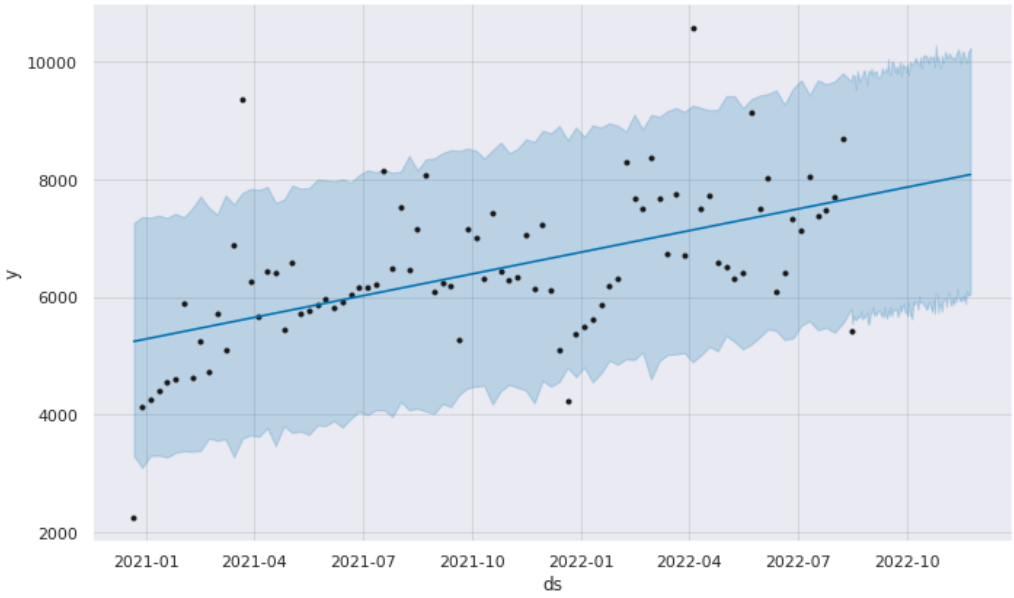
Hemos decidido tomar tres algoritmos para nuestra predicción:

- Regresión Lineal
- Decision Tree Regressor
- FB Prohpet

Métricas finales del modelo

Luego de haber realizado los tres modelos pudimos obtener los siguientes resultados:

- El Decision Tree Regressor nos dio un score del 10% por lo cual lo descartamos para su uso
- La regresión lineal nos dio un accuracy del 78%
- Prophet nos entrega una predicción de órdenes con valor promedio, máximo y mínimo para cada iteración el cual se aproxima mucho a la realidad.



Conclusiones

En base al modelo Prophet pudimos obtener los valores de ordenes por cada día a futuro con los valores posibles mínimos, máximos y medios.

Consideramos que es importante que este modelo se siga alimentando de nueva información como así también de nuevas variables para que la brecha entre los 3 resultados sea cada vez menor y así llegar a un nivel de predicción casi exacto.