Mejorando la eficiencia en la entrega de alimentos: Análisis predictivo de tiempos de entrega en una aplicación de Food delivery

CoderHouse, Data Science

Vila Navarro María Paz

Tabla de contenidos:

Abstract con objetivo y contexto comercial

Este proyecto analiza datos de una aplicación de Food delivery para mejorar la eficiencia en la entrega de alimentos y la experiencia del cliente. Los datos incluyen información sobre las ordenes y los repartidores. El objetivo es identificar patrones y tendencias en el desempeño de los repartidores para mejorar la calidad del servicio y aumentar la satisfacción del cliente.

El sector de entrega de alimentos ha experimentado un gran crecimiento y las aplicaciones de delivery se han vuelto populares. A medida que la competencia aumenta, las empresas deben identificar formas de mejorar la eficiencia y la calidad del servicio. La clientela se ha empezado a acostumbrar a este tipo de servicio y para poder sobresalir ya no alcanza con ofrecerlo, sino con que calidad se hace. Poder entregar los alimentos en buen estado y más rápido que la competencia se vuelve crucial.

El objetivo de este proyecto es desarrollar un modelo de análisis predictivo que permita estimar el tiempo de entrega de un pedido en función de diferentes variables, tales como la distancia, el repartidor asignado y otras características relevantes del pedido.

Trabajo en el equipo de operaciones de la empresa por lo que tengo mucha información y acceso a la data reciente sobre como se han desarrollado las metricas.Es responsabilidad de nuestro equipo mejorar los tiempos de entrega.

Se espera que este proyecto tenga implicaciones importantes para la empresa de entrega de alimentos y el sector de delivery en general. Al mejorar la eficiencia y la calidad del servicio, la empresa puede aumentar la satisfacción del cliente y mantenerse competitiva en un mercado en constante evolución.

Contexto analítico

Hemos obtenido de nuestra base de datos todas las ordenes realizadas en el último tiempo.Los datos están etiquetados y tengo la intención de predecir el tiempo necesario para entregar una orden de acuerdo a las condiciones de esta, por lo que planeo utilizar modelos de regresión para abordar este problema de aprendizaje supervisado.

Preguntas/hipótesis

Pregunta: ¿Qué características de las ordenes y repartidores son más relevantes para el tiempo de una orden?

Hipótesis:Creo que impactan mas la distancia, condiciones climáticas, y el tráfico.

Pregunta: ¿Cómo varía el tiempo total de entrega en base a la distancia recorrida?

Hipótesis: A más distancia las entregas duran más tiempo

Pregunta: ¿Cómo varía el tiempo total de entrega en base a las condiciones climáticas?

Hipótesis:A mejores condiciones climáticas menor tiempo de entrega

Pregunta: ¿Cómo varía el tiempo total de entrega en base a el tráfico?

Hipótesis: A mayor tráfico mayor tiempo de entrega

Pregunta: ¿Cómo varía el tiempo total de entrega en base a el rating y la edad?

Hipótesis: A mayor rating menor tiempo y a mas edad mas tiempo.

Descripción de los datos

El dataset es un conjunto de datos descargados del sitio Kaggle.

Food delivery refiere a un servicio de mensajería en el que un restaurante, una tienda o una empresa independiente de entrega de alimentos entrega alimentos a un cliente. Por lo general, un pedido se realiza a través del sitio web o la aplicación móvil de un restaurante o tienda de comestibles, o a través de una empresa de pedidos de alimentos. Los artículos entregados pueden incluir entradas, guarniciones, bebidas, postres o artículos comestibles y, por lo general, se entregan en cajas o bolsas. El repartidor normalmente conducirá un automóvil, pero en las ciudades más grandes donde las casas y los restaurantes están más cerca, pueden usar bicicletas o motos.

Es un dataset público,y se puede acceder en el siguiente Link.

A continuación se provee de una breve descripción de las variables:

- **ID:** identificador unico
- **Delivery_person_ID:** identificador del repartidor
- **Delivery_person_Age:** edad del repartidor
- **Delivery_person_Ratings:** puntuación del repartidor
- **Restaurant_latitude:** latitud restaurante
- **Restaurant_longitude:** longitud restaurante
- **Delivery_location_latitude:** lat punto de delivery

- **Delivery_location_longitude:** long punto de delivery
- **Order_Date:** fecha de la orden
- **Time_Orderd:** hora de creación de la orden
- **Time_Order_picked:** hora de recogida de la orden
- **Weatherconditions:** condiciones del clima
- **Road_traffic_density:** densidad del trafico
- **Vehicle_condition:** condiciones del vehiculo en int
- **Type_of_order:** tipo de orden (ej snack,meal)
- **Type_of_vehicle:** tipo de vehiculo
- **multiple_deliveries:** si la orden era bundle es decir si tiene mas de un punto de entrega que puede afectar al tiempo
- **Festival:** tipo de festividad o no
- **City:** tipo de aglomerado
- **Time_taken(min):** tiempo total de entrega

El dataset contiene **20** variables con **45593** registros. Consulté por valores duplicados pero no tiene.En un principio tuvo valores nulos.

Resumen de valores nulos en decimales

ID	0.000000
Delivery_person_ID	0.000000
Delivery_person_Age	0.040664
Delivery_person_Ratings	0.041849
Restaurant_latitude	0.000000
Restaurant_longitude	0.000000
Delivery_location_latitude	0.000000
Delivery_location_longitude	0.000000
Order_Date	0.000000
Time_Orderd	0.037966
Time_Order_picked	0.000000
Weatherconditions	0.013511
Road_traffic_density	0.013182
Vehicle_condition	0.000000
Type_of_order	0.000000
Type_of_vehicle	0.000000
multiple_deliveries	0.021780
Festival	0.005001

Decisiones respecto de los valores nulos luego del EDA:

- Variables cuantitativas: El rating y la edad se los reemplazó por la media
- Variables cualitativas: se eliminó los nulos

Feature selection: como parte de este cree nuevas variables que puedan aportar al analisis

- **Distancia:** es la diferencia en kilómetros entre la latitud y longitud del restaurant y la latitud y longitud del punto de entrega. Aspecto que creo que puede impactar mucho en el tiempo de entrega.
- **Densidad_Trafico_Clima:** combino las variables "Weatherconditions" y "Road_traffic_density".Representará las condiciones generales en que se desarrolló cada orden
- **Velocidad_promedio**:utilizando la variable que creada distancia y el tiempo total de la orden. Esta variable creo que puede ser particularmente importante para el modelo, reflejará la eficiencia del repartidor en función de la velocidad
- Order year: creada con order date.
- Order month:creada con order date.
- Order day:creada con order date.

Transformación de las variables: para poder proceder al análisis tuve que transformar variables.

- **Time_taken:** elimine las letras y convertí a números flotantes.
- **Rating**: convertí a número flotante
- **Edad**: convertí a numero flotante
- **Time orderd**: convierto a objetos timedelta. Y luego a números enteros que serían la cantidad de segundos transcurridos desde la medianoche.
- **Time order picked:**convierto a objetos timedelta. Y luego a números enteros que serían la cantidad de segundos transcurridos desde la medianoche.

Eliminación de variables:

En el primer relevamiento de las variables, nos encontramos con algunas que no aportan información adicional por lo que se las excluyó. Las variables excluidas son:

- **ID**:es el identificador único de la orden
- **Delivery_person_id:** es el identificador único del repartidor
- **Order date**: es la fecha total, la eliminé porque cree las de year,month,day.
- **Las dos columnas de latitud:** las eliminé porque solas no aportan, las utilicé para crear las distancias.
- **Las dos columnas de longitud**:las eliminé porque solas no aportan, las utilicé para crear las distancias.

Así que entre las eliminadas y las añadidas tengo 19 columnas.

EDA: Exploratory Data analysis

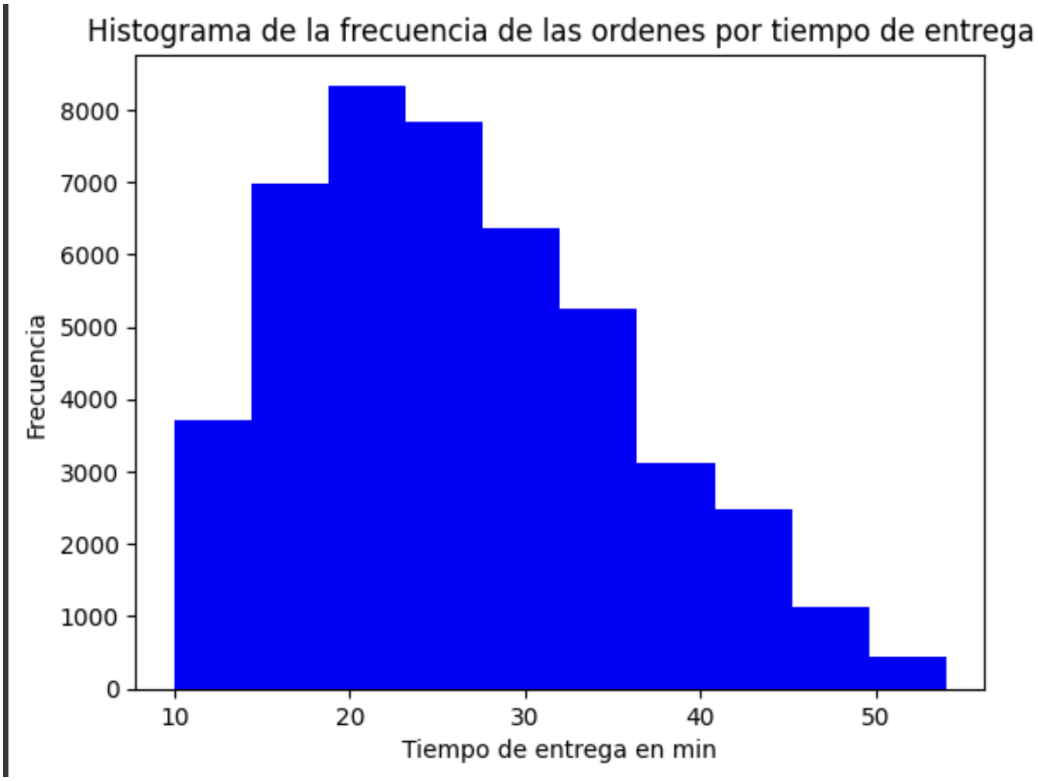
El total de las 19 variables son :

- 8 variables categóricas
- 8 variables numéricas (cuantitativas)
- Time delta:3

Análisis univariado

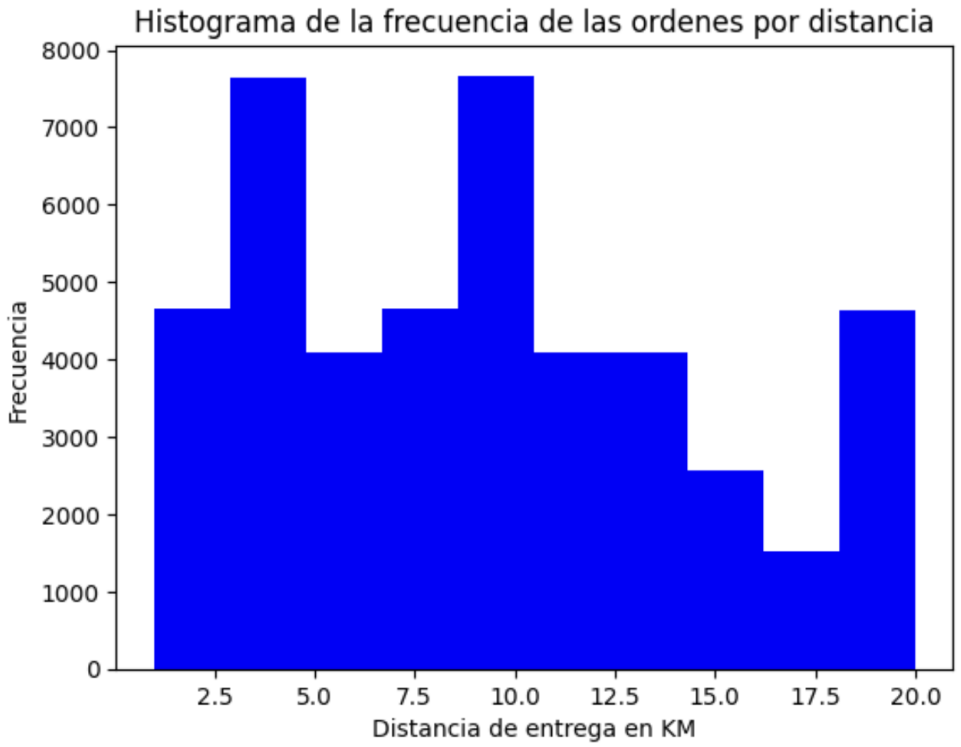
El EDA lo comencé por las variables numéricas.

Describí la variable cuantitativa continua, **Time taken,** a partir de sus medidas de tendencia central y dispersión, y con un gráfico univariado de tiempos.En resumen, tiene una media, mediana y moda muy similares (cerca de 26 minutos) Pero al hacer deep dive, el gráfico muestra una distribución asimétrica.



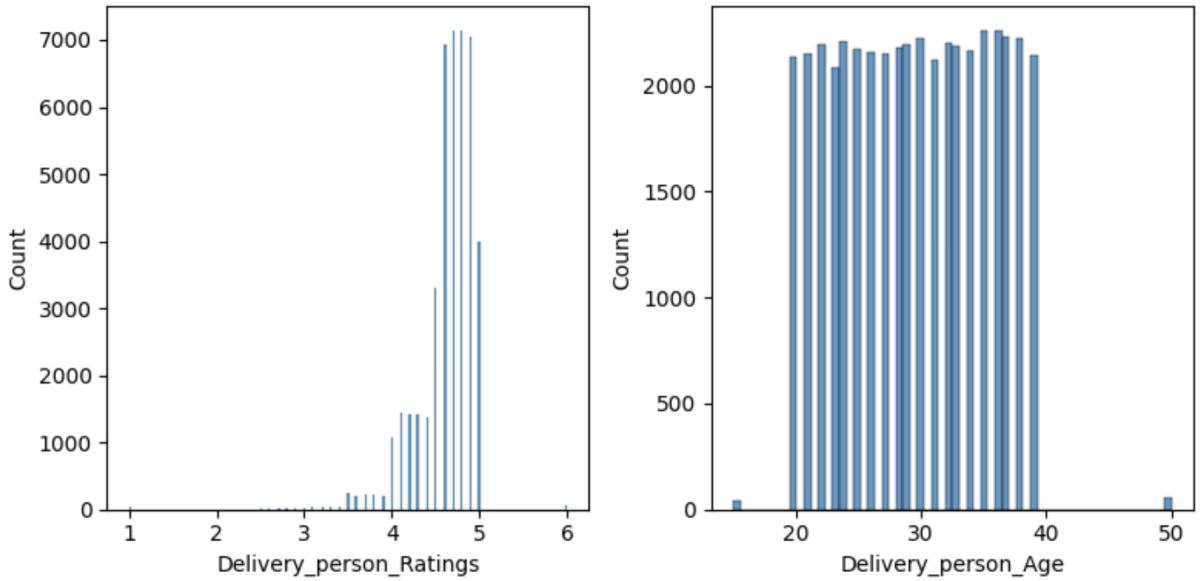
Miré los mismos datos de la **distancia**. Esta es una variable cuantitativa continua. A partir de las medidas veo que la mediana y la media son iguales (9) pero la moda un tanto menor (7). Los valores parecen ser dispares. El gráfico muestra que es asimétrica y por momentos

Uniforme, hay muchos valores que se parecen entre sí, con 3 picos en distancias cercanas a 3,9,19)



Respecto al **rating y la edad** puedo decir que son dos variables cuantitativas continuas. A partir de sus histogramas puedo ver que el rating tiene una distribución asimétrica con una media de 4.63. Con mucha concentración entre 4.5 y 4.9. Lo que muestra que en general los repartidores tienen buena performance.

La edad tiene una distribución uniforme entre 15 años y 50 (como outliers).Con una media de 29 años, pero cómo es uniforme puedo ver que desde los 20 hasta cerca de los 40 años hay una similar cantidad de repartidores por edad.



Respecto a las variables cualitativas utilicé los gráficos de torta para describirlas.

Weatherconditions: se distribuyen de una forma demasiado simétrica, casi que tenemos el mismo porcentaje para todos los climas.

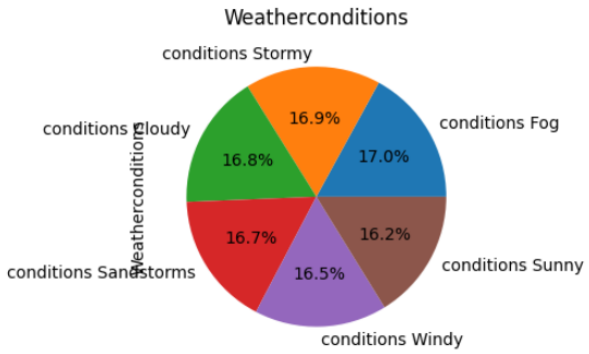
Type of order: ocurre lo mismo que en el clima

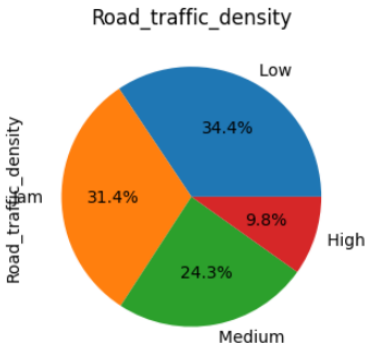
City: veo que la mayor cantidad de órdenes se dieron en áreas metropolitanas, en segundo lugar en áreas urbanas, y muy poco (0.4%) en Semi urbanas.

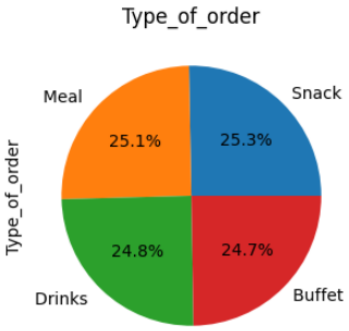
Road traffic density: El mayor porcentaje de órdenes en Low, luego en jam, luego en medium y luego en high. así que puedo pensar que es una ciudad con no tanto tráfico.

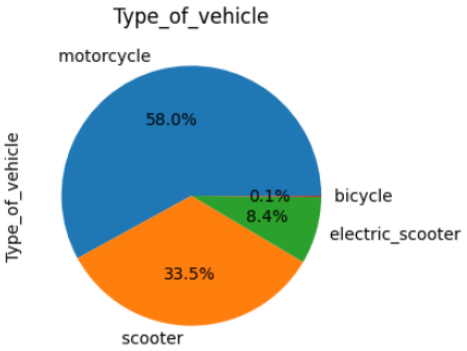
Type of vehicle: ampliamente se han utilizado motos (58%) para entregar las órdenes, luego scooters y luego scooters eléctricos y bicicletas.

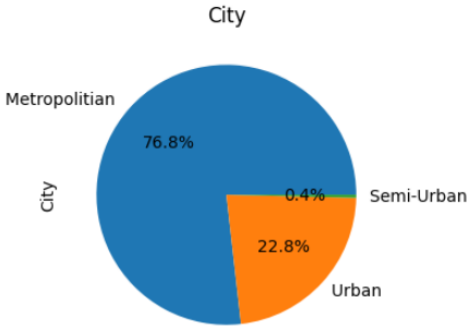
Festival: el 98% de las órdenes no se hizo durante un dia festivo

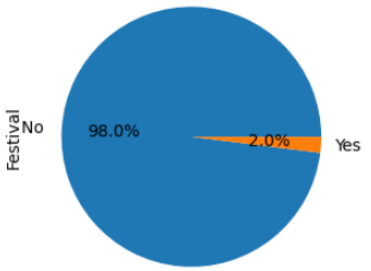






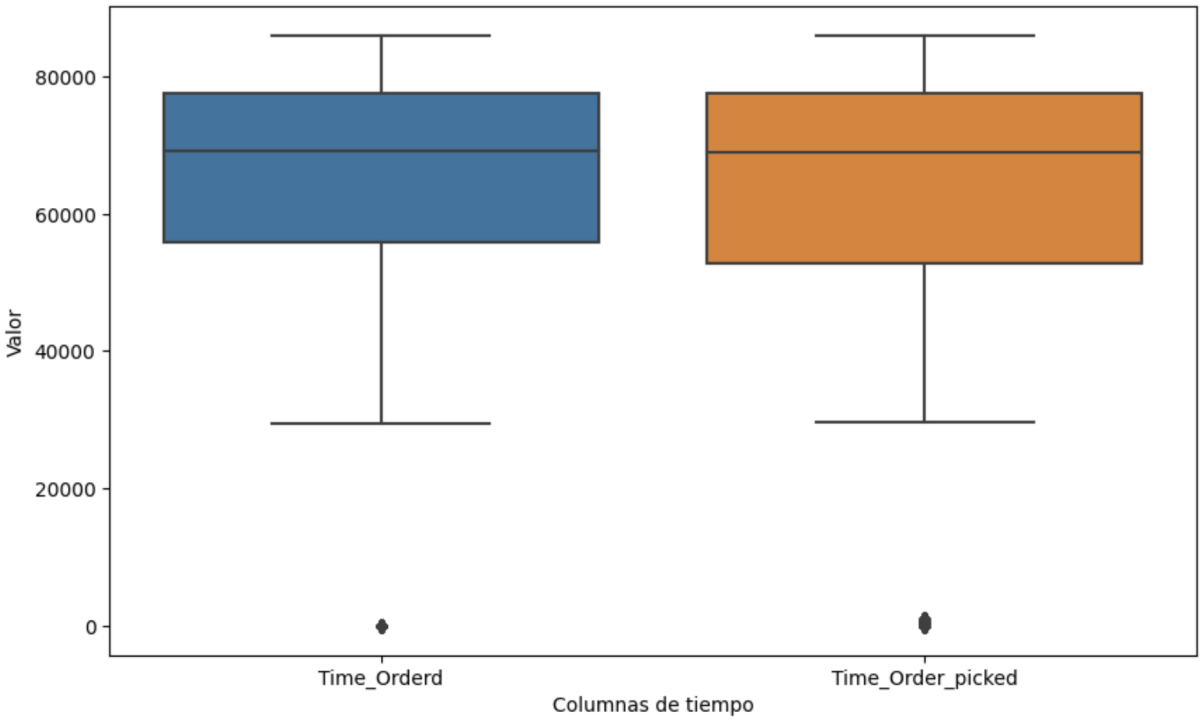






Outliers

Al analizar las columnas de tiempos encontré outliers . 404 para Time orderd y 1178 para time order picked.Decido finalmente conservar estas filas ya que puede ser que las ordenes hayan sido creadas o recogidas en distintos horarios y no afecta el analisis.

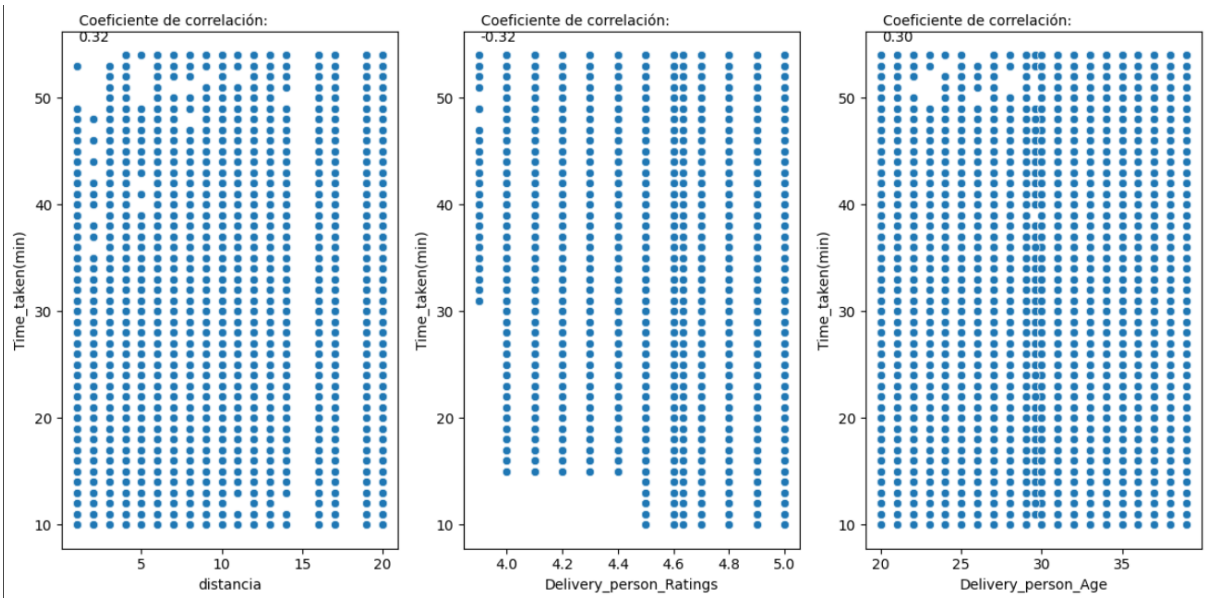


Al buscar los outliers del resto de las variables me llamaron la atención los del rating, los calculé y eran 1099. Decidí cambiarlos por la media ya que los valores, tal como vimos, están concentrados y prefiero no perder esos registros.

Análisis bivariado

Del análisis bivariado, no se observaron correlaciones que permitieran escribir una variable en función de otra.

Relación de variables cuantitativas con el tiempo



En los tres gráficos se observa mucha correlación.

En el de la relación con la distancia al tener un coeficiente de correlación de 0.32 puedo pensar que hay una muy leve correlación positiva entre las dos variables, por lo que podría decir que hay una tendencia a que a medida que aumenta la distancia aumente el tiempo, pero es muy baja esa tendencia.

Esto no contradice mi hipótesis de que la distancia era uno de los factores que mas impactaba el tiempo de las órdenes. Que a mas distancia mas tiempo, pero tampoco es muy consistente porque muestra una correlación muy baja

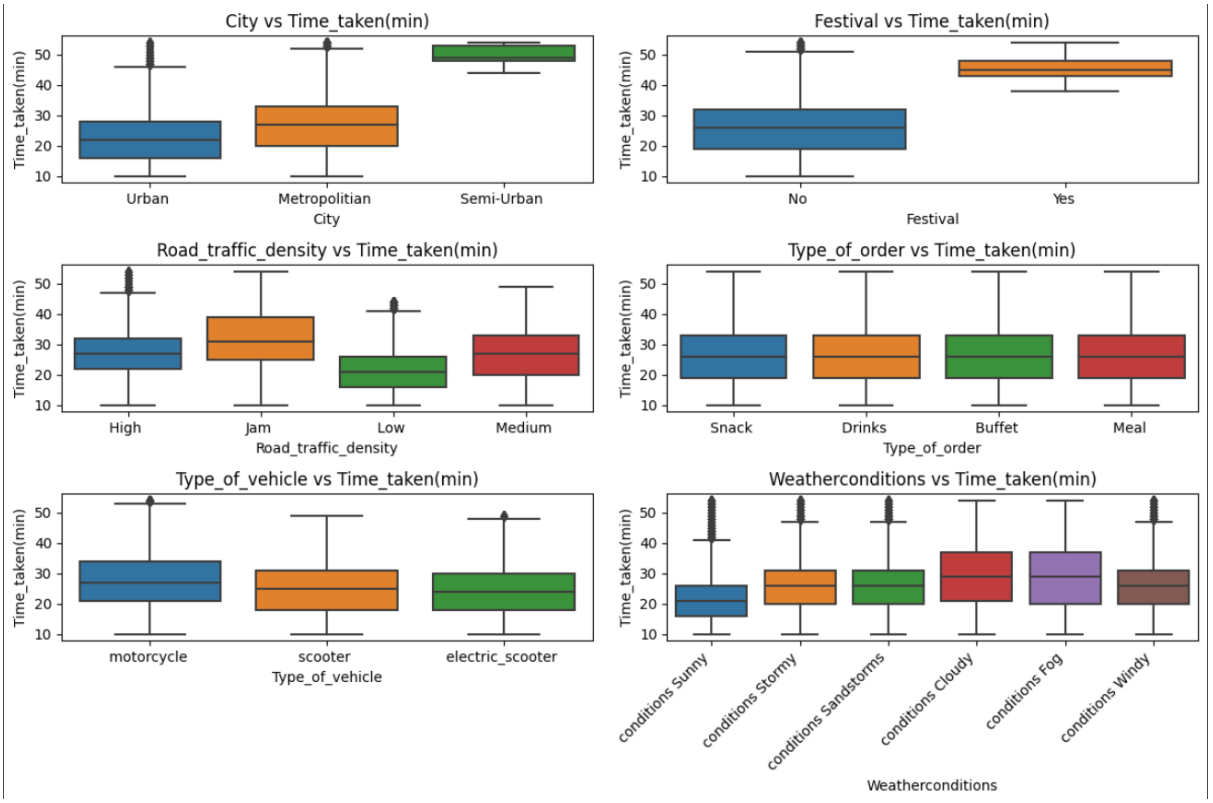
En el de la relación con el rating del repartidor tiene un coeficiente negativo y veo que no hay mucha relación entre ambas. hay una leve correlación negativa, por lo que hay una tendencia a que cuando el tiempo aumenta el rating baja.

Esto no contradice mi hipótesis de que el rating impacta al tiempo de las órdenes. Que a menor rating mas tiempo, pero tampoco es muy consistente porque muestra una correlación muy baja

En el de la relación con la edad muestra una correlación similar al de la distancia, correlación muy leve positiva, por lo que hay una leve tendencia a que cuando aumenta la edad tarden mas tiempo.

Esto no contradice mi hipótesis de que la edad impacta al tiempo de las órdenes. Que a mayor mas tiempo, pero tampoco es muy consistente porque muestra una correlación muy baja

Respecto a la relación con variables cualitativas:



Relacion entre city y tiempo: el tipo de ciudad impacta al tiempo. Veo que en areas semi urbanas el tiempo es mayor,mientras en las urbana es menor pero con mayor distribucion de los valores, la distancia entre los bigotes es mayor. En lareas metropolitanas es donde la distancia entre bigotes es mayores, es decir hay mas amplitud de valores de tiempo, y una media mayor que en urbanas pero menor que en semi urbanas.

Relacion entre festival y tiempo puedo decir que cambia mucho el tiempo en base a si hay un festivo o no. cuando es yes, el tiempo aumenta.

Relacion entre traffic density y tiempo el tráfico afecta al tiempo. Cuando es low veo que los tiempos son menores y cuando es jam son mayores. Tambien tienen distintas medias en base al trafico

Relacion entre type of order y tiempo veo que no afecta al tiempo. Casi que sin importar el tipo de orden se tarda lo mismo en entregarse, el tiempo se comporta de forma muy similar.

Relacion entre type of vehicle y tiempo sucede algo similar que con type of order pero veo que las cajas de scooter y electric scooter muestran menos tiempo. por lo que podria decir que las motos y bicis ddemoran un poco mas aunque no tiene gran impacto esta variable.

Relacion entre weather conditions y tiempo veo que el clima afecta al tiempo. en condiciones cloudy o fog se tiende a tardar mas, aunque tambien tienen mayor amplitud sus

cajas, por lo que puede suceder que que se tarde menos de lo pensado. En condiciones sunny es cuando menos se tarda
Luego de los analisis puedo ver que mi suposición de que time taken iba a estar mas correlacionada con la distancia, la edad y el rating es efectivamente asi, estas variables tienen mas correlación en los dos primeros casos positiva y en el tercero negativa . Aunque esto no muestra fuerte correlación, veo que para el resto de variables hay menos.

También es interesante la correlación positiva entre time ordered y distancia.Y con time order picked.

Feature Engineering y Forward selection

Antes de poder aplicar los modelos de Machine Learning se transformaron las variables:

1. Codifiqué mis variable categóricas con Label encoder.
2. Eliminé las columnas ID y Delivery_person_ID
3. Convertí las variables de hora a objetos timedelta y las convierto a números enteros que serían la cantidad de segundos transcurridos desde la medianoche
4. Elimino columnas Order_Date, las de latitud y longitud ya que fueron utilizadas para otras cosas y no cumplen más función.
5. Luego de aplicar forward selection elimino Order_year y order_day ya que me demuestra que no aportan valor

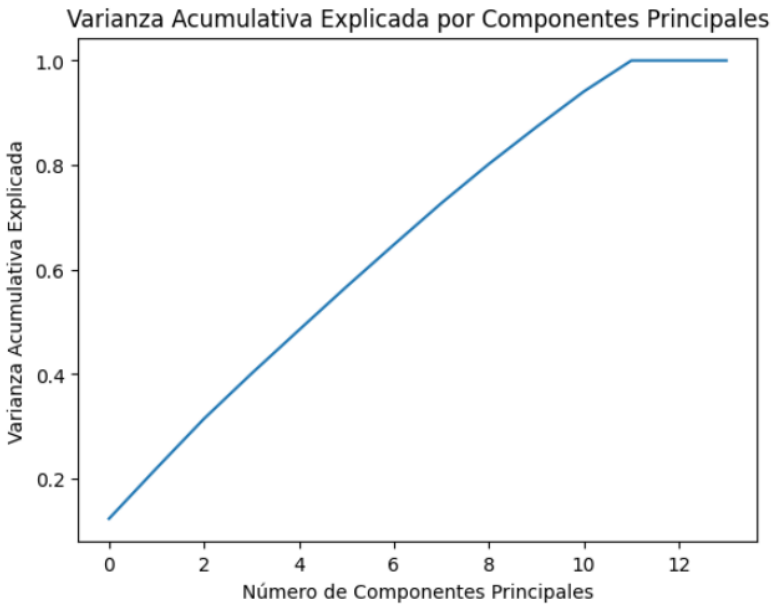
Algoritmo

Como el objetivo de este proyecto es desarrollar un modelo de análisis predictivo que permita estimar el tiempo de entrega de un pedido en función de diferentes variables, utilicé primero un **algoritmo de regresión lineal** para entrenar con los datos que tengo.

Para esto se utilizó sklearn con un modelo de regresión lineal, con train y test, con un test de 30% del dataset.

Se evaluaron dos métricas, el MAE y el R2.El MAE fue igual a 5.35, si bien está en la misma unidad de mi variable (minutos) preferí utilizar el r2 ya que es más claro su resultado. El valor del R2 fue igual a 0.48, con este valor no podía decir que era un buen modelo, en el 50% de los casos predecirá bien y en el otro 50% predecirá mal.Por otro lado, no estaba ni overfitiado ni underfitiado, no tenía esos problemas.

Para mejorar el modelo utilicé el **Análisis de Componentes Principales, como técnica de reducción de dimensionalidad.** Con las 16 variables que lo hice, cada una aportó menos del 12% de explicación de la varianza. Pero en general son extremadamente pequeños los porcentajes (cercanos a cero), por lo que parece que no aportan mucha información. 13 de ellas aportaron muy poca información, y 3 nada.Lo que se comprueba en el siguiente gráfico:



Con esta información elegí entrenar un modelo de **Random forest** con 10 variables (84.6% de la varianza) y evaluar que tal funcionaba. Pase de 16 variables a 10 (mi reducción). Pero el resultado fue malo, el accuracy fue de 6%.

A partir de este resultado se decidió **crear nuevas variables** para mejorar el modelo:

1. Cree la nueva variable**"Densidad_Trafico_Clima**" al combinar las variables "Weatherconditions" y "Road_traffic_density".
    a. Representará las condiciones generales en que se desarrolló cada orden.
2. Tambien creé "**Velocidad_promedio**" utilizando la variable que ya había creado (distancia) y el tiempo total de la orden.
    a. Esta variable creo que puede ser particularmente importante para el modelo,refleja la eficiencia del repartidor en función de la velocidad

Luego, volví a implementar un **PCA** para ver el impacto de las nuevas variables al predecir el Time_taken. Separé en train y test y normalice los datos para que funcione mejor. Los resultados , que reflejan las proporciones de varianza explicada por cada componente principal, mejoraron pero aún siguen siendo pequeños. El primer componente explica el 16% mientras los dos últimos nada.

Al evaluar las 16 variables en base al resultado del PCA, observo que para llegar al 90% debo quedarme con 11 variables. Con esa cantidad de variables entrené un **Random Forest.** El resultado reflejó un mejora del modelo pero sigue siendo bajo,un accuracy del 7%.

Para mejorar la precisión de mi modelo predictivo apliqué un modelo de ensamble, utilicé **XGBoost - Regresión.**El resultado del error cuadrático medio fue de 1.47 (**MSE**), es decir, en promedio las predicciones del modelo tienen ese error cuadrático. Cuanto menor sea su valor, mejor será el rendimiento del modelo. Sin embargo, lo que se considera "bueno" o "malo" puede variar según la naturaleza del problema y los datos por lo que haré una **Cross Validation para evaluar y mejorar la capacidad del modelo de generalización.**

Para esto preprocese los datos, cree mi modelo de regresión usando un **Random Forest Regressor, y utilicé la función Cross_val_score de Scikit-Learn para hacer la validación cruzada.**El resultado fue de un valor promedio de R^2 de 0.78 con una desviación estándar de 0.02. Esto significa que el modelo de regresión RandomForestRegressor tiene un buen rendimiento en promedio, y es una mejora respecto al anterior donde obtenía un R2 de 0.48.

Sin embargo, es importante tener en cuenta que la desviación estándar de 0.02 también es relativamente baja, lo que sugiere que el rendimiento del modelo es consistente en los diferentes pliegues de la validación cruzada.

Por lo que puedo ver que las nuevas variables e incorporaciones implicaron una mejora y me quedo con ellas y con este modelo como final.