Delivery Time Predictions

Introduction

Our company manages the delivery of grocery orders placed online. An important element of optimising drivers’ working time is predicting how long it will take to deliver the order to the customer's address. Our system collects GPS data from the driver’s devices then the partitioning algorithm splits this data into segments, identifies all stops and relates them to orders delivered by the driver. We use this information to predict the duration of future deliveries. The current prediction is really simple - it calculates the mean from all data collected so far and uses it as a predicted delivery time for every future order.

Our long-term goal is to have the best possible predictions, so that predicted and actual times do not differ much. This will allow us and other teams from external systems to plan with better accuracy.

Data Definition

Data of historical deliveries in our system is stored in 4 separate tables defined as:

1. orders:
   - order_id - unique identifier of an order (primary key),
   - customer_id - identifier of a customer that made an order,
   - sector_id - identifier of a geographic area,
   - planned_delivery_duration - estimated delivery duration in seconds,
2. products:
   - product_id - unique identifier of a product (primary key),
   - weight - weight of a product in grams,
3. orders_products
   - order_id - identifier of an order,
   - product_id - identifier of a product,
   - quantity - number of these products in an order,
4. route_segments:
   - segment_id - unique identifier of a segment (primary key),
   - driver_id - identifier of a driver,
   - segment_type - type of a segment (DRIVE, STOP),
   - order_id - identifier of an order delivered in the segment (only when type = STOP),
   - segment_start_time - timestamp of segment start,
   - segment_end_time - timestamp of segment end,

Task

One day the product manager gives you a task:

> *We predict delivery times for future orders just by calculating the mean from all collected data. We are currently trying different solutions. Recently we discovered that delivering to single-family houses takes significantly less time than delivering same size orders to apartment buildings. Unfortunately, we don't have data about building type, but maybe you will find any similar trends or correlations in the data we have? Then, the developers team could translate your findings into improvements in the prediction algorithm.*

Instructions

You can use any tools you find suitable (relational DB system, Looker Studio, spreadsheet, etc.).

Part 1. Data modelling [SQL]

1. You received a droptime.sql file. Use it to create a database and populate it with data. You can use any SQL-compliant relational database engine. The input file was prepared for MySQL, using other DB systems may require subtle changes to the queries.
2. Create a query which will return data about the total weight per product ordered by the customer with customerId = 32 delivered on February 13, 2024. The query should return data in the following schema:

productId	totalWeight

   sorted by totalWeight ascending.

Export results to CSV file *name_surname.csv* and include it as part of your final solution.

Part 2. Data analysis and visualisation

Create a report in which you will include charts and briefly present your line of thinking. Write down assumptions you made, describe your methodology, e.g. how you filtered data etc. The report should be easy to understand for non-technical people. Name your report *name_surname_analysis.pdf* and include it as part of your final solution.

Don't make any assumptions about data quality. The data may contain some erroneous samples or corrupted values.

1. Generate a histogram showing the actual delivery length with 1 minute granularity (rounded up).
2. Generate a histogram showing prediction error (difference between planned and actual delivery times).

3. We received insight from our drivers that delivering in one of the sectors is significantly longer than in other sectors. Generate a chart to visualise this hypothesis.
4. Play with the data by grouping, aggregating and remodelling it. Are you able to find any correlations or trends that could be valuable for prediction quality improvement? Describe briefly your findings and visualise them on charts.

Part 3. Building and verifying the hypothesis

Create a report in which you will briefly present your line of thinking. Write down assumptions you made, describe your methodology, e.g. how you filtered data etc. The report should be easy to understand for non-technical people. Name your report *name_surname_research.pdf* and include it as part of your final solution.

Don't make any assumptions about data quality.

1. The current prediction algorithm is very naive. It calculates the mean from all collected data and applies it to every future order. We need to explore alternative ideas. One of them is predicting delivery times per sector. Describe how you would validate this hypothesis using available data.
2. Using the data, propose some alternative method/algorithm that will predict delivery times more accurately. Describe the methodology to validate the new algorithm.
3. Why could some deliveries take more time? For example, some buildings don't have elevators etc. Describe your ideas.
4. What additional data would be worth collecting for future analysis of this domain?
5. What is the risk of over- or under-estimating the delivery times?