unacast.

Unacast Hands-on challenge

Naveen Kumar Bhageradhi

Tasks

- Analyze venue visitation patterns.
- Handle data inconsistencies.
- Create a scalable pipeline for data pre-processing and forecasting total visitor count for each venue (multiple time series forecasting).

unacast.

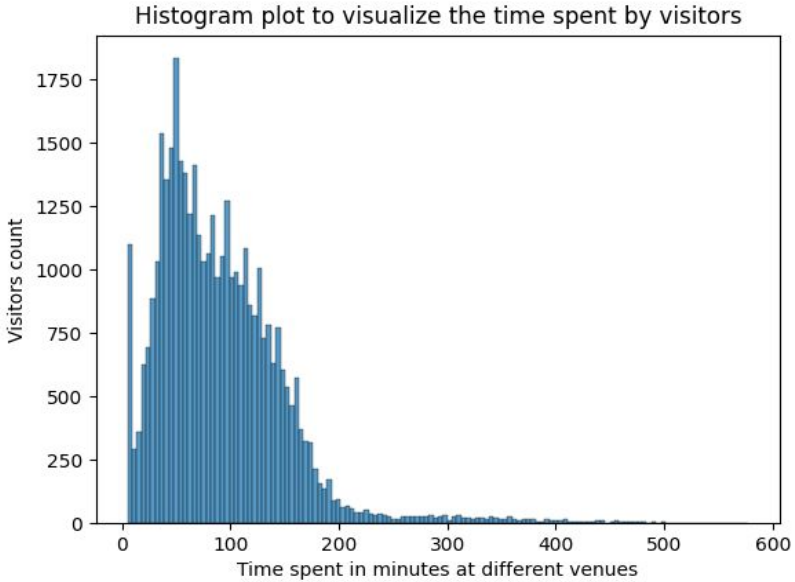
Data pre-processing

- The dataset includes daily visitor information across **15** venues, categorized into **3** groups (UN, CO CI).
- Remove **duplicate** rows (total duplicates: **198**).
- Handle **venue type** column (missing values: **92**, unknown: **38**).
- Verify if data is available for **each venue.**

unacast.

Data pre-processing

- Handle **visit_end_time** column (missing values: **396**).
- **Median** time spent by the visitor at venue: **81** min



Data pre-processing

- **Coffee_place** and **university** venues have more visitors on **weekdays**, while **cinema** venues have slightly higher visitor numbers on **weekends**.
- Different venues under same category have **similar** visitation patterns.

unacast.

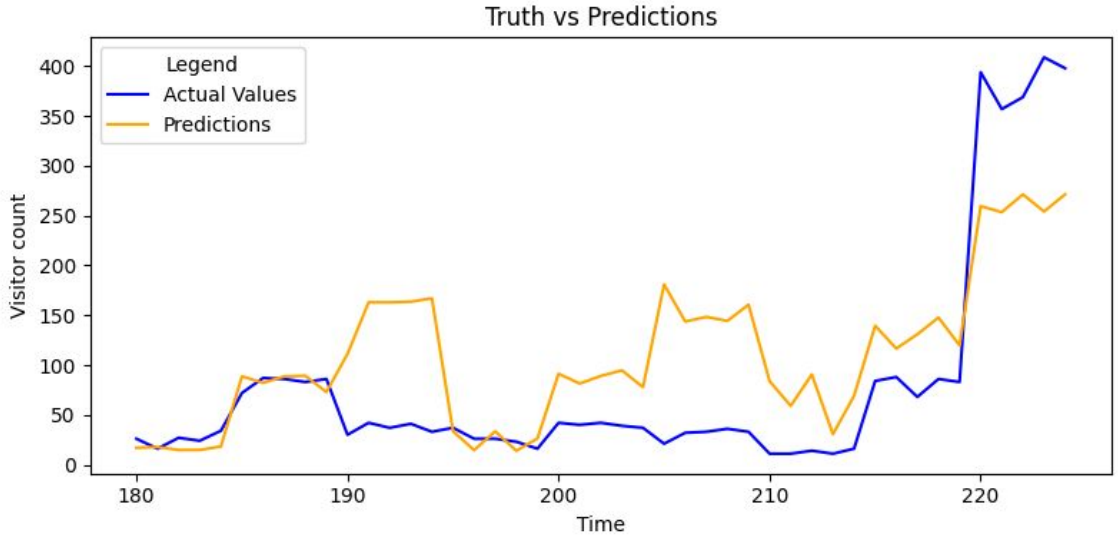
Lightgbm model

- Exclusive Feature Bundling (EFB) method is useful for multiple time series forecasting.
- **Training and test data**: files from October 28 to November 11
- **Daily forecasting data**: files from November 12 to 17
- **Features:** visitor lag values from 1 to 6, day_of_week, day, month, year, venue_id
- **Parameters:**
  - ‘**colsample_bytree**': 0.3, '**learning_rate**': 0.05, '**max_depth**': 5, '**n_estimators**': 50, '**num_leaves**': 10
  - These parameters are selected based on grid search cv

unacast.

Results

- Mean Absolute Error (**MAE**) on test data: 60.17



unacast.

For more robust model

- Develop advanced features such as **expanding mean** from lag variables, integrating **weather data**, and encoding **holidays** or **festivals** to enhance temporal understanding for each venue.
- On larger datasets train **diverse models** representing different forecasting method families and perform a comparative evaluation of their results.

unacast.

Visitation forecasting pipeline

- For **new data** on each day first model is **updated** with the actual visitor count and then visitor count is **forecasted** for next day.