Chicago Divvy Bike-sharing program bike usage prediction

This project aims to predict the daily usage counts of the Chicago Divvy bike-sharing program.

Hang Tian

Mengqi Li

Vaishnavi Meka

Introduction

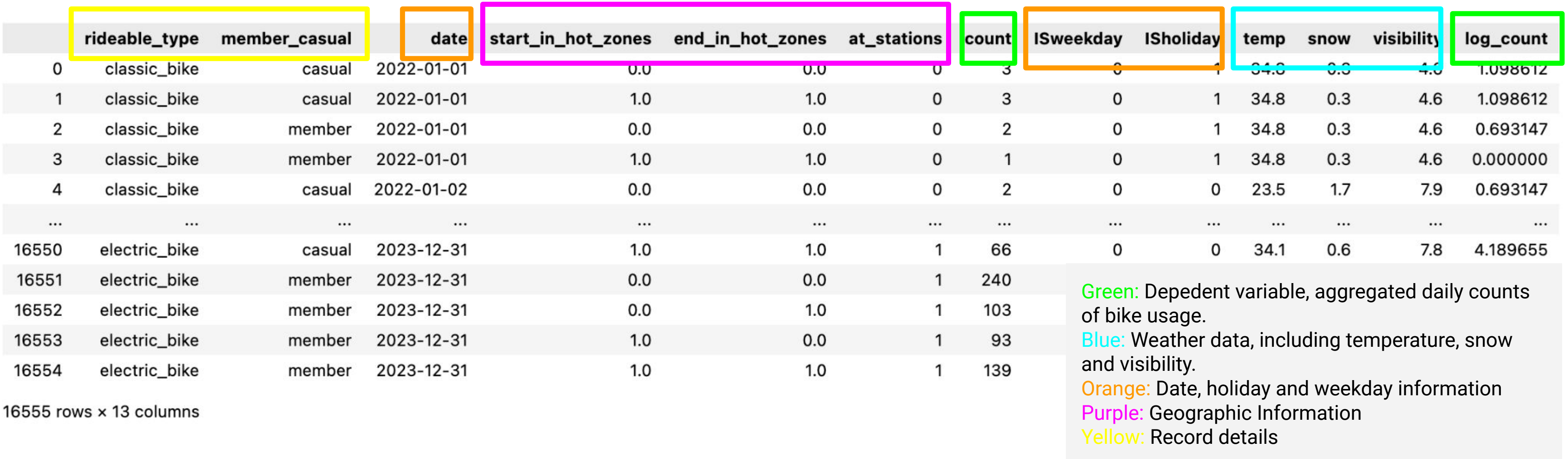
Domain Problem

- The Divvy bike-sharing program in Chicago is a popular transportation option for both residents and visitors. Understanding the factors that drive bike usage can help the program's stakeholders make informed decisions about station placement, bike fleet management, and service optimization. This project will analyze historical data to develop a predictive model for bike usage.

Stakeholders

- Our main stakeholder is the Chicago Department of Transportation (CDOT). Our predicted bike usage information can not only guide the planning and allocation of shared bike resources, but also support futural decisions about where to expand or adjust services to integrate with other modes of transportation. To summarize, our project could help optimize the city’s bike-sharing infrastructure.

Data



	rideable_type	member_casual	date	start_in_hot_zones	end_in_hot_zones	at_stations	count	ISweekday	ISholiday	temp	snow	visibility	log_count
0	classic_bike	casual	2022-01-01	0.0	0.0	0	3	0	1	34.8	0.3	4.6	1.098612
1	classic_bike	casual	2022-01-01	1.0	1.0	0	3	0	1	34.8	0.3	4.6	1.098612
2	classic_bike	member	2022-01-01	0.0	0.0	0	2	0	1	34.8	0.3	4.6	0.693147
3	classic_bike	member	2022-01-01	1.0	1.0	0	1	0	1	34.8	0.3	4.6	0.000000
4	classic_bike	casual	2022-01-02	0.0	0.0	0	2	0	0	23.5	1.7	7.9	0.693147
...	...	...	...	...	...	...	...	...	...	...	...	...	...
16550	electric_bike	casual	2023-12-31	1.0	1.0	1	66	0	0	34.1	0.6	7.8	4.189655
16551	electric_bike	member	2023-12-31	0.0	0.0	1	240						
16552	electric_bike	member	2023-12-31	0.0	1.0	1	103						
16553	electric_bike	member	2023-12-31	1.0	0.0	1	93						
16554	electric_bike	member	2023-12-31	1.0	1.0	1	139						

16555 rows × 13 columns

Green: Depedent variable, aggregated daily counts of bike usage.
Blue: Weather data, including temperature, snow and visibility.
Orange: Date, holiday and weekday information
Purple: Geographic Information
Yellow: Record details

1. Bike-sharing trip data: The Divvy bike-sharing program is ran by CDOT (Chicago Department of Transportation), this dataset is highly reliable. We used two-year bike trip data from 2022 to 2023 for this project.
2. Chicago neighbourhoods shapefile: This dataset comes from Chicago open data portal, it's well maintained by Chicago government and highly reliable .
3. Holiday information: According the calendar, we've compiled all the weekdays and holidays from 2022 to 2023.
4. Daily weather data: Weather data for 2022 to 2023 collected from the Visual Crossing website.

Methods

Data Preprocessing

- **Aggregation:** Final model was aggregated group by features within records then joined with data from other sources.
- **Geographic Information Extraction:** Counting the total start and end points within each neighborhoods of Chicago city, the top five with highest number are labeled as hot zones.
- **Outlier removal:** There are certain aggregated records with few or several times more records than others, these daily counts are identified as outliers

Model Selection

- LinearRegression is the initial model we used.
- Random Forest model to get the features importance.
- KNN model
- Decision Tree model
- SVM model
- XGboost model gives the best results
- Neural Network model

Results

1

R-squared: 0.97

The final model achieved an R-squared value of 0.97, indicating most of the variations can be explained by the features. It is quite high because we divided daily counts into many groups.

2

RMSE of Logged count: 0.3

The rooted mean squared error (RMSE) of the logged model was 0.3, indicating the average predicted bike usage counts varies from 0.74 - 1.34 times of the true value. Still needs improvement.

3

Key Findings

According to feature importances in random forest model, bike type, member type, holidays were key categorical features. Temperature was found the key numerical feature.

Next Steps

Work remaining

Lower RMSE

More features, especially numerical features that have influence on bike usage.

More records, including more years.

Limitation

Geographic Infromation Combination

More useful geographic zones needed to solve stakeholders' need.

How to combine geographic information and machine learning is still in debate.