Bike Sharing Analysis and Modelling

Lawrence Rosen

Executive Summary

Objective:

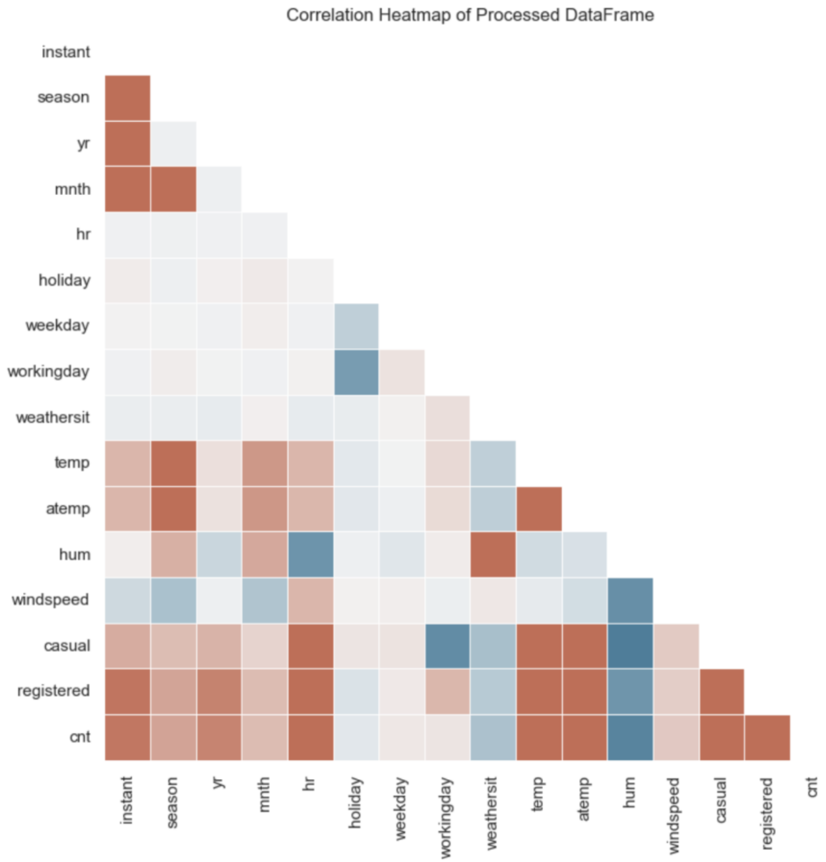
- Find relationships between time, weather and other data on the demand of bike rentals.
- Create predictions for a 4-week period based on these features including engineered features.

Key Findings:

- Hour of day has the greatest effect on the rentals.
- Ratio of registered to casual is less on weekends and also less in non-working hours.
- Baseline model with no feature engineering achieved:
- Predictive modelling (LightGBM) for rentals achieved an RMSE of 0.58 after feature engineering, recursive elimination of features and model hyperparameter tuning.
- Holiday data seemed to not have an effect.
- Casual users more influenced by weather and weekends — opportunity to target them differently with promotions.

Recommendations:

- More analysis into registered and casual users. Promotions and other data could offer more insight into any anomalies and improve accuracy.
- Inclusion of new features such as price of rental, payday, alternative transportation cost, fuel costs, vehicle taxes, congestion charges i.e. ULEZ enforcement in London etc.
- Investigation of the quality of some features. Holidays didn't seem to have an effect - maybe this data is wrong.
- Location data of rentals – planning how many bikes in each location – can calculate missed revenue



Data Overview

Dataset Details:

- Columns: 17, Rows: 17379
- No Missing Data

-

dteday – Date

season – Season
1: Spring
2: Summer
3: Fall
4: Winter

yr – Year
0: 2011
1: 2012

mnth – Month (1 to 12)

hr – Hour (0 to 23)

holiday – Whether the day is a holiday or not

day_of_week – Day of the week

workingday – If the day is neither weekend nor holiday = 1, otherwise = 0

weathersit – Weather situation
1: Clear, Few clouds, Partly cloudy
2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds
3: Light Snow, Light Rain + Thunderstorm + Scattered clouds
4: Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog

temp – Normalised temperature in Celsius (divided by 41)

atemp – Normalised feeling temperature in Celsius (divided by 50)

hum – Normalised humidity (divided by 100)

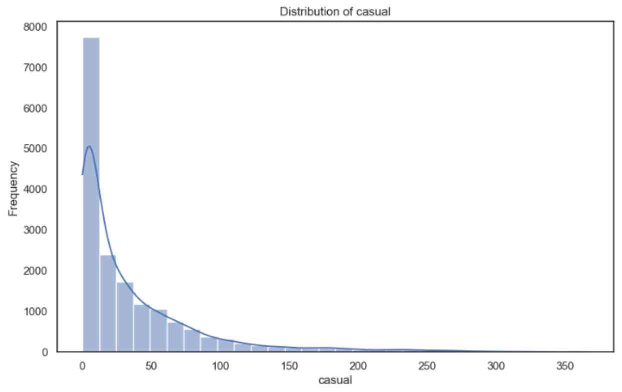
windspeed – Normalised wind speed (divided by 67)

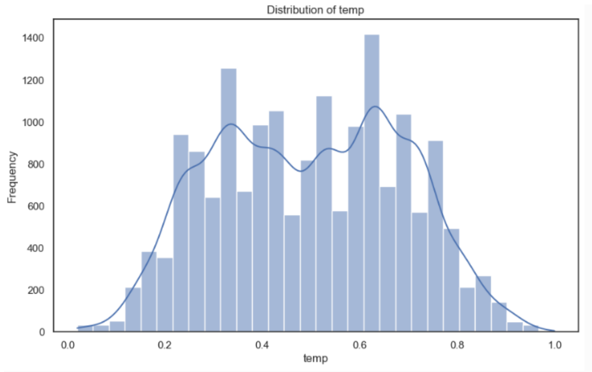
casual – Count of casual users

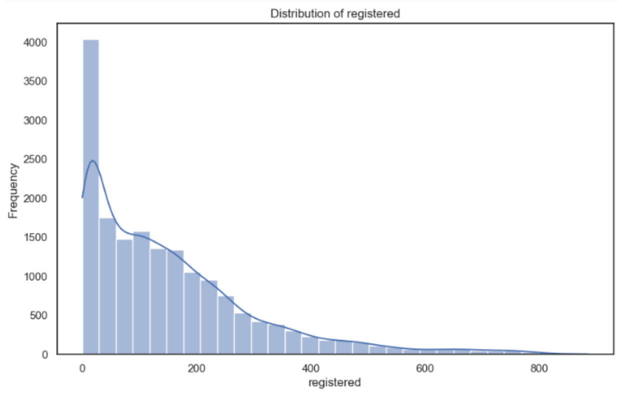
registered – Count of registered users

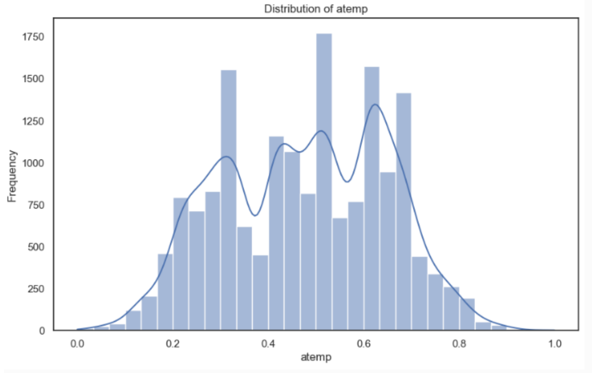
```
season: [1 2 3 4]
yr: [0 1]
mnth: [ 1  2  3  4  5  6  7  8  9 10 11 12]
hr: [ 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23]
holiday: [0 1]
weekday: [6 0 1 2 3 4 5]
workingday: [0 1]
weathersit: [1 2 3 4]
```

Data Overview

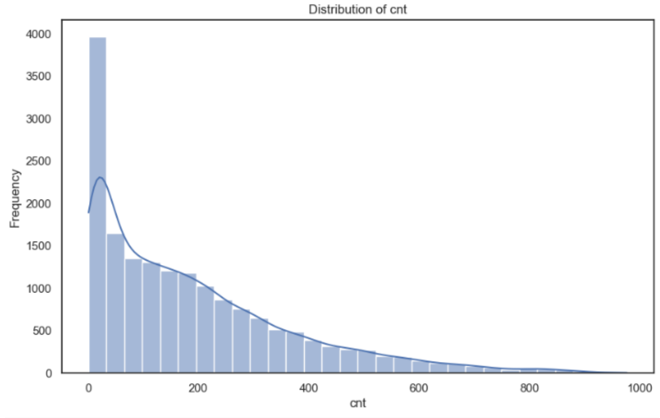


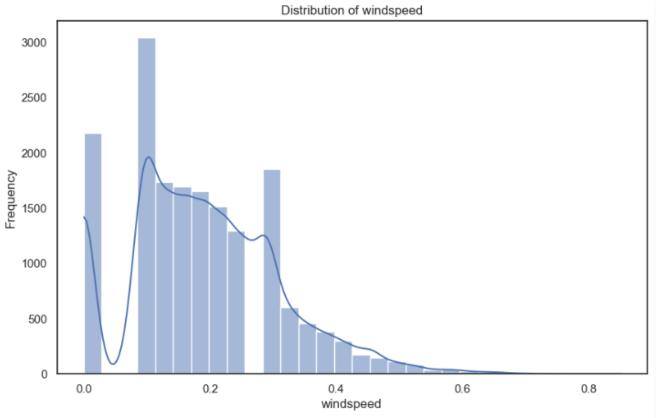


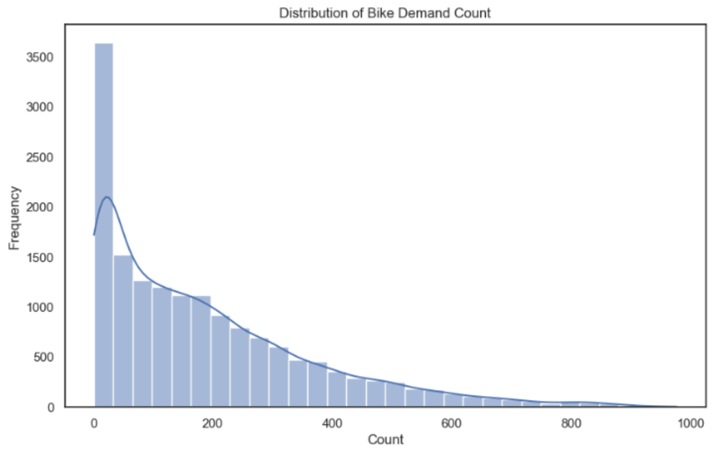


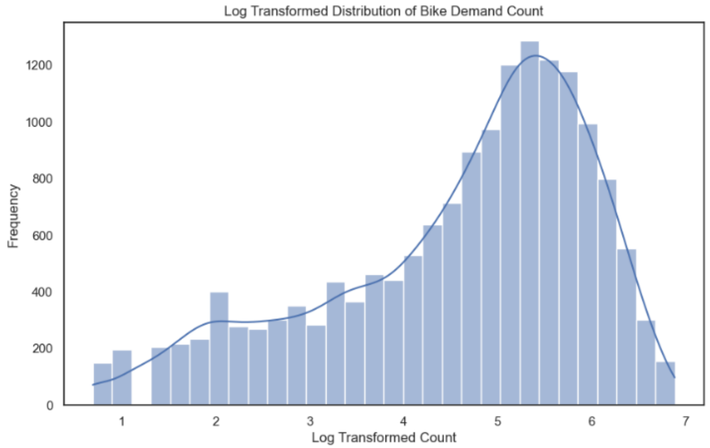


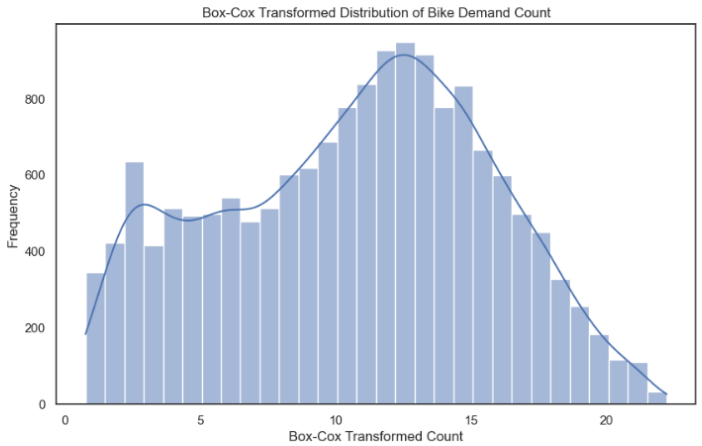
- Count, Registered and Casual are ‘right skewed’
- Weather features are normalised - windspeed is a bit skewed.
- Windspeed also appears to miss some data – possibly the sensitivity of the measuring equipment?









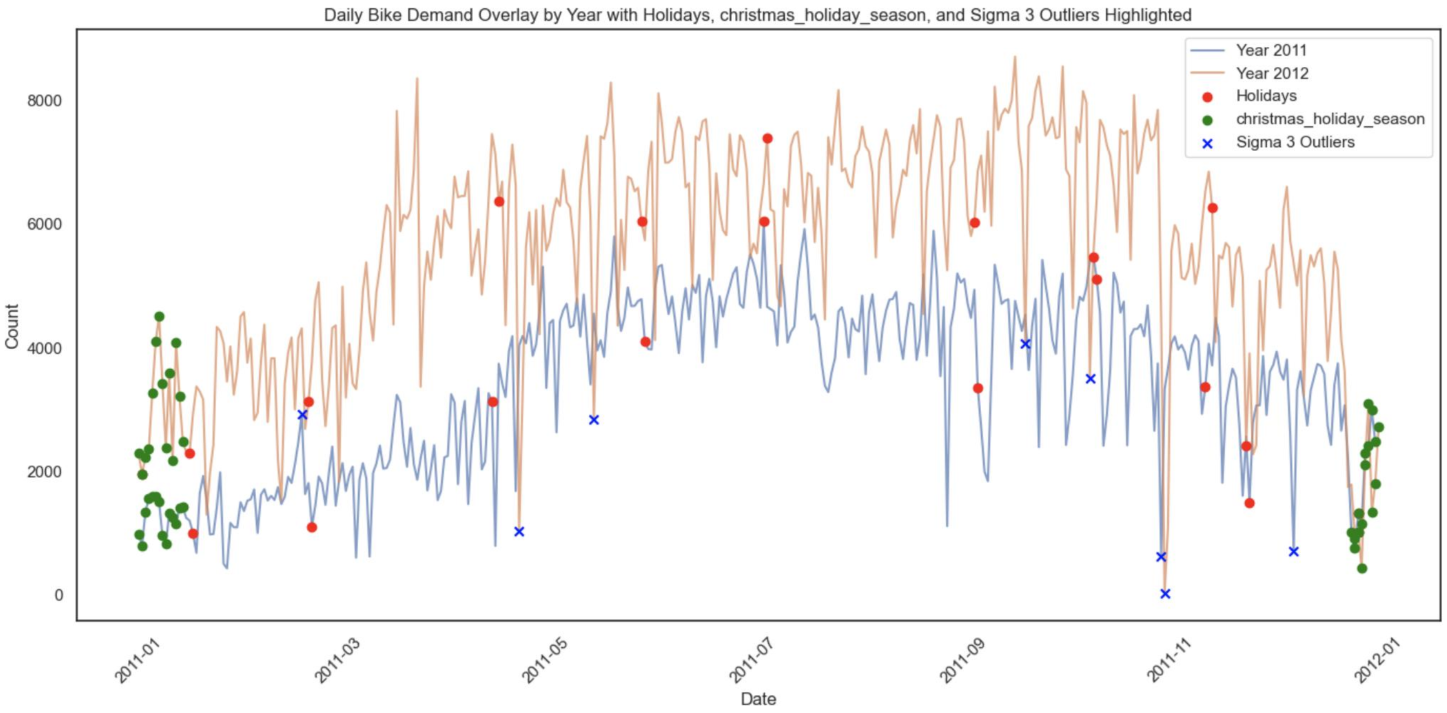


Target Variable - Count

- The target count was ‘right skewed’
- Many models, particularly linear regression, perform better with normally distributed data.
- A logarithmic transformation is often used but I decided to use a box cox as it approximated normality better.
- Normality is not crucial for tree-based models, but it can still be helpful for model interpretability and for testing other regressors.

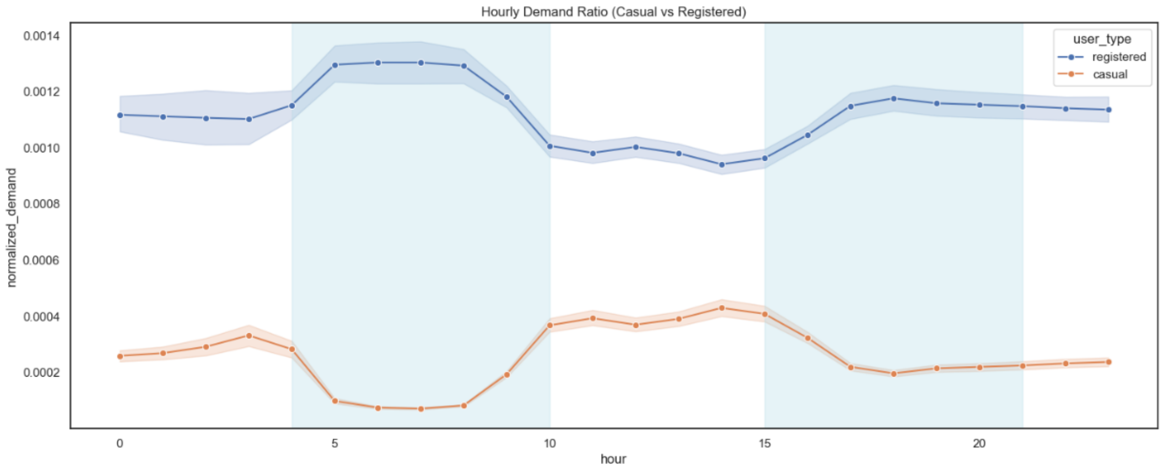
Yearly Profile

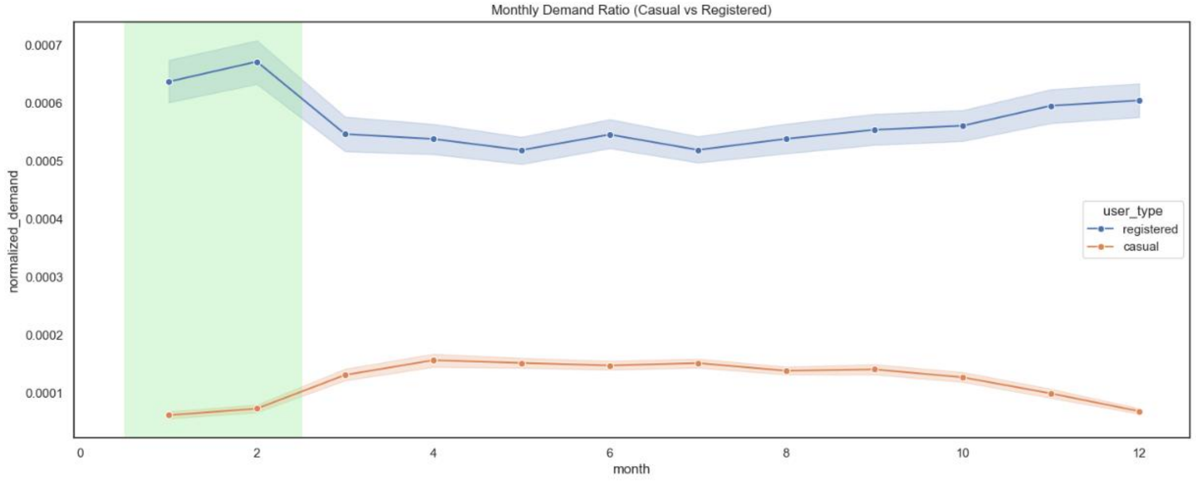
- Yearly Seasonality
- Increase over hotter parts of year.
- Large effects from Christmas
- Holidays do not seem to line up with any peaks
- A few potential anomalies/big swings
  - April large swing – possible data logged on wrong day?)
  - November Two outliers align, maybe planned outage? Or a holiday which isn't marked?

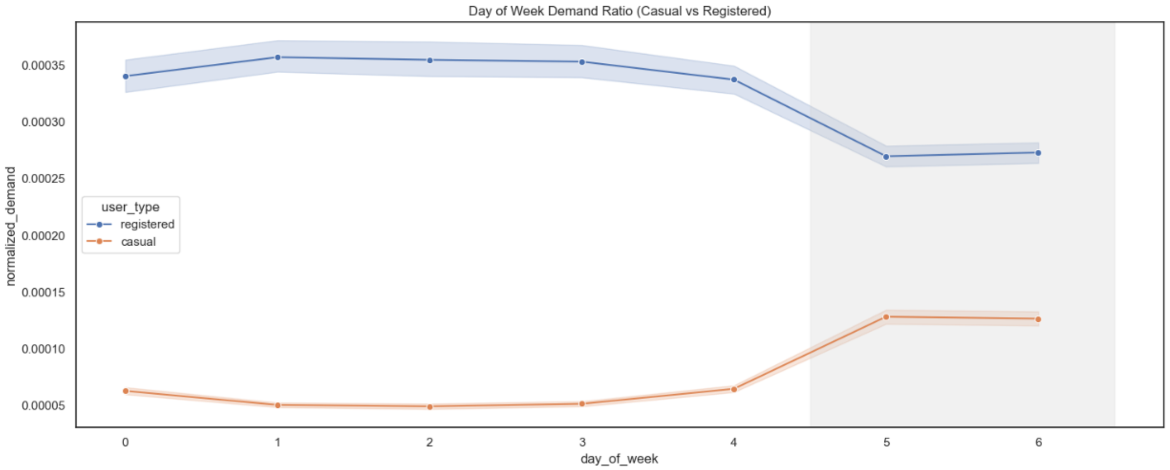


More Time Profiles

- The ratio of registered to casual users increases in colder periods of the year.
- This suggests casual users are more dependent on weather and registered users are more dependent on commuting features.
- is also demonstrated in the rush hour periods and weekend vs kday comparisons.

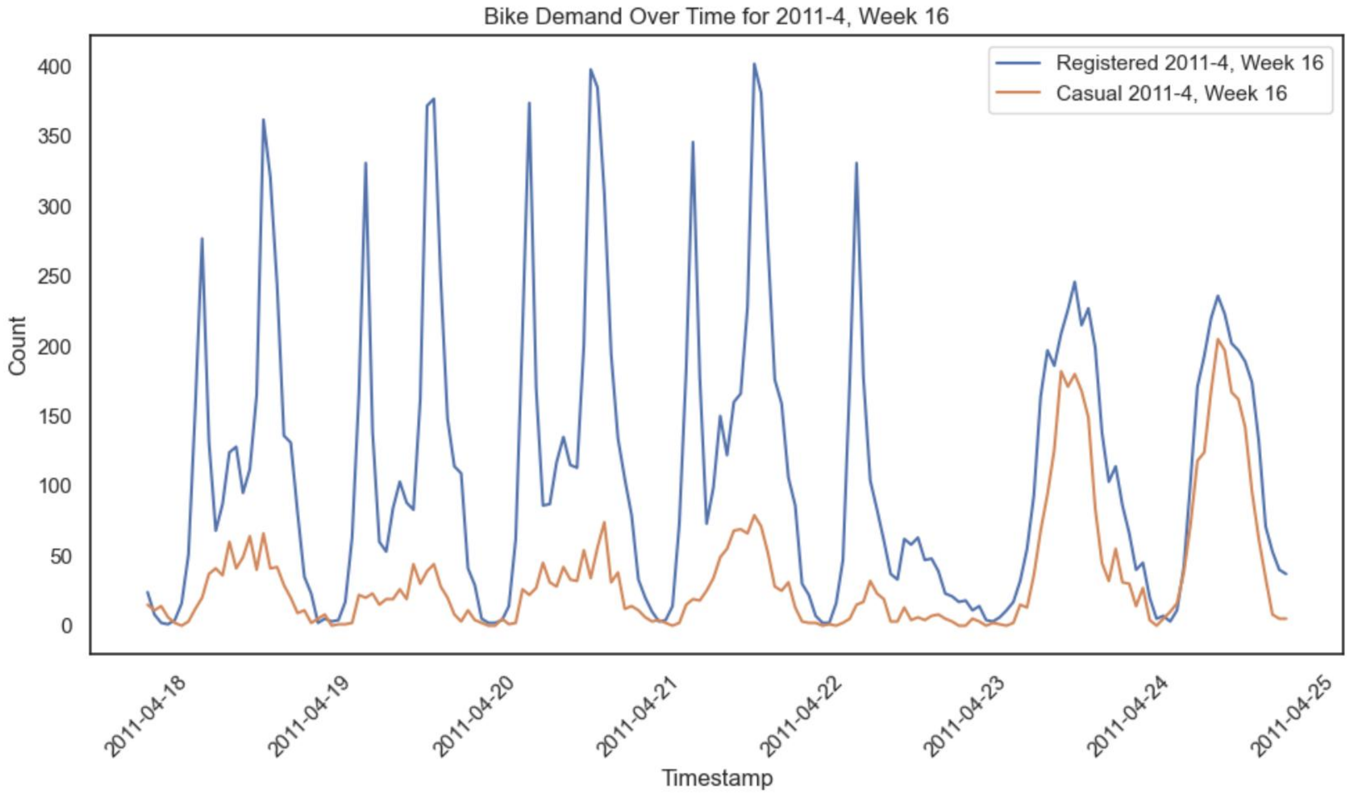


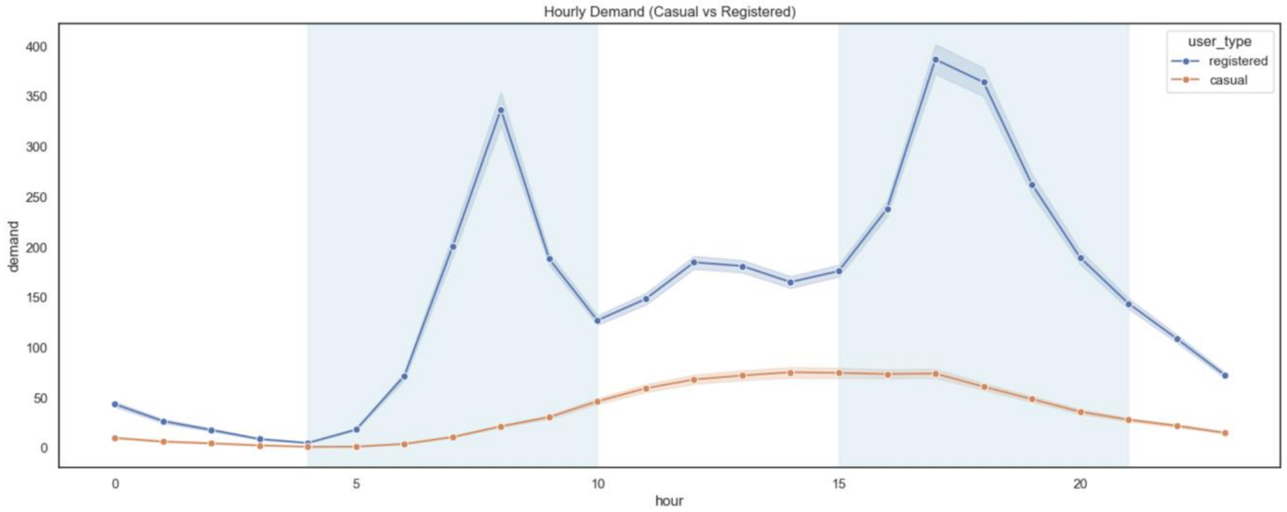




Hourly Profile of Collisions

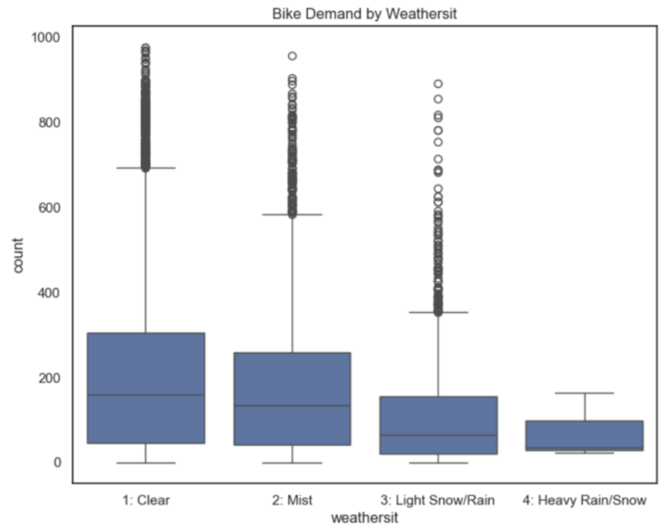
- Clear trend for working days with two peaks around rush hours.
- Saturday and Sunday more smoothed and less travel.
- Higher Peak for afternoon maybe work fatigue.
- It would be interesting to see before and after covid difference in profiles due to work from home.

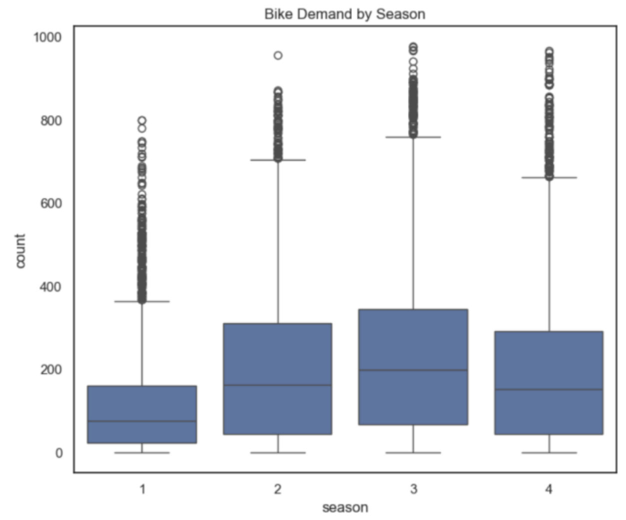




Weather and Season

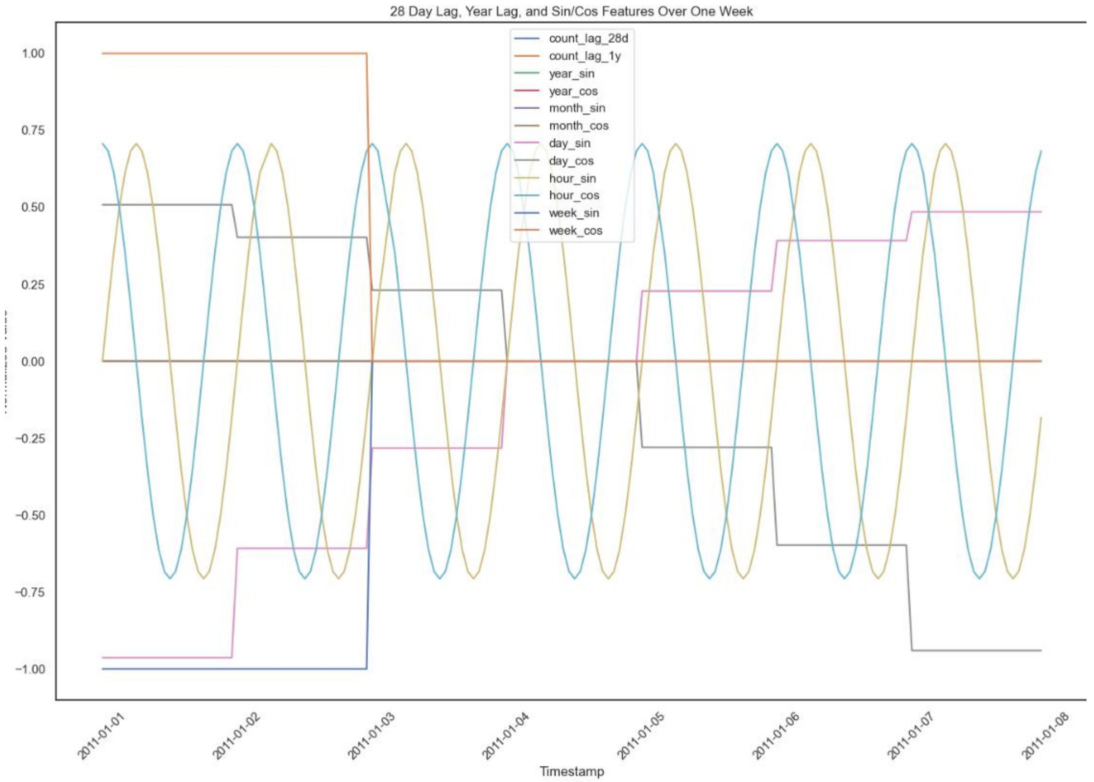
- Warmer Seasons = greater demand
- Nicer weather = greater demand





Feature Engineering

- Although machine learning models can inherently figure out relationships between features, it has been shown in research that it is better to combine features to give the models help.
- Features were created using combinations of weather.
- Ratio of registered user features were also created for different time splits.
- Any features that could cause data leakage were only made using the train data such as the registered and casual ratios.
- Registered and casual columns were removed as this is equivalent to the amount of rentals each day.



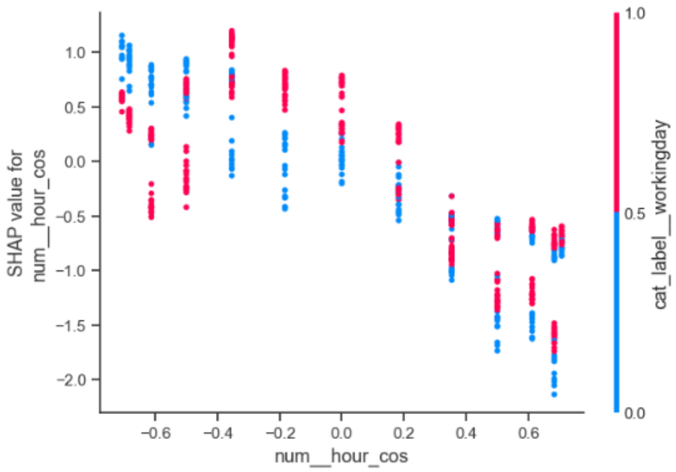
Shap analysis

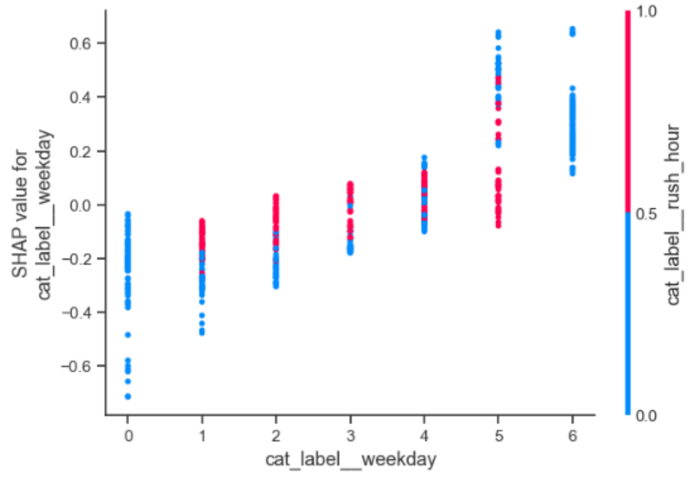
SHAP values quantify each feature's contribution to the prediction of a specific instance.

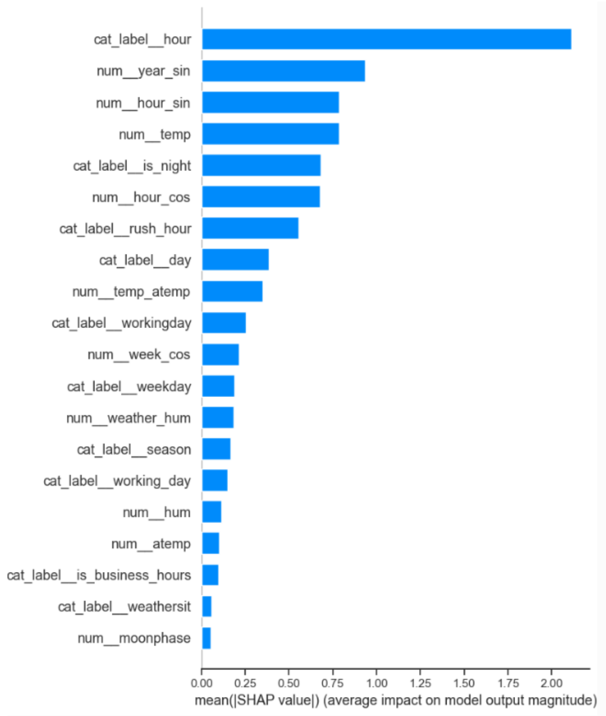
Why Use SHAP for Feature Importances?

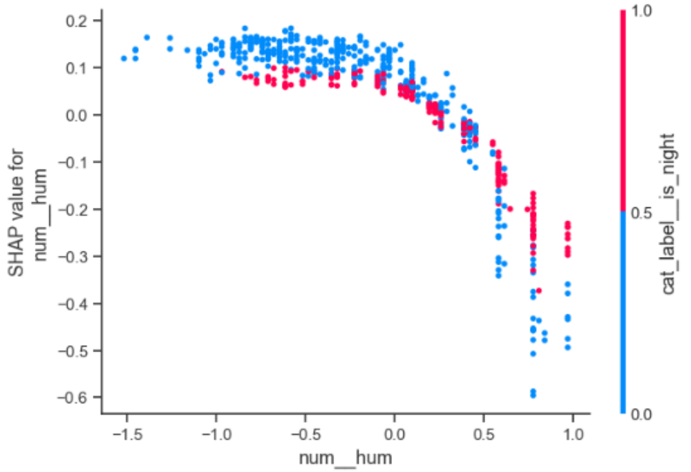
Accurate and Consistent: SHAP provides a consistent method of attributing feature importance, satisfying desirable properties like local accuracy and consistency.

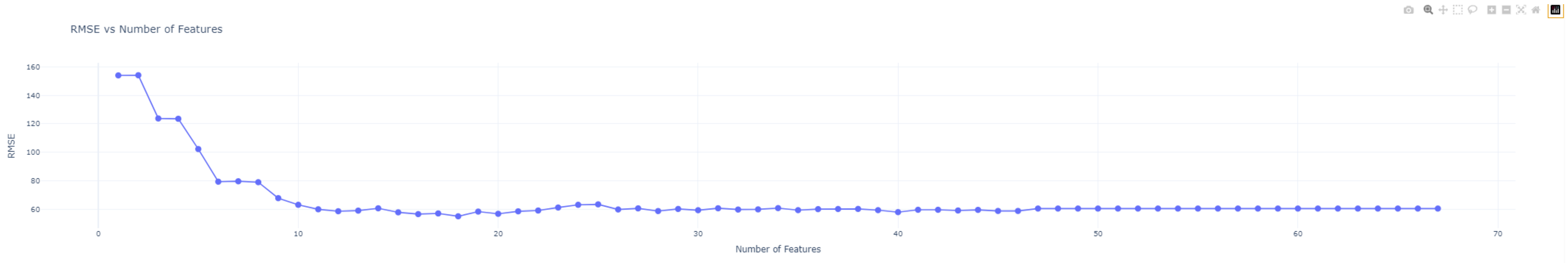
- **Model-Agnostic**: Works with any machine learning model, making it versatile for different algorithms (e.g., tree-based models, neural networks).
- **Interpretability**: SHAP offers clear, interpretable insights into how individual features influence predictions, providing a deeper understanding of the model.
- **Global and Local Explanations**: Not only shows overall feature importance but also allows for individual prediction insights, aiding in better decision-making.











Feature Reduction with SHAP

- **Why Use SHAP for Recursive Feature Elimination (RFE)?**

- **Improved Feature Selection**: Using SHAP values for RFE allows for the identification of the most important features based on their actual contribution to the model, rather than relying on statistical measures alone.

- **More Reliable Results**: SHAP-based RFE offers a more robust and data-driven approach to feature elimination compared to traditional methods, accounting for interactions between features.

- **Enhanced Model Performance**: By removing irrelevant or less important features, SHAP RFE can help reduce overfitting, improve model interpretability, and increase overall performance.

Hyperparameter Optimisations

A Hyperameter optimiser called Optuna was used to test different model parameters for the LightGBM model using a 5 fold time series cross validation strategy.

Selected Features

n_estimators = 200

- The number of boosting iterations (trees) in the model.
- Higher values can lead to better accuracy but may cause overfitting.

learning_rate = 0.08

- Controls the contribution of each tree to the final model.
- Lower values make the model more robust but require more estimators.

max_depth = 4

- The maximum depth of each tree.
- A lower value prevents the model from becoming overly complex and overfitting.

num_leaves = 90

- The maximum number of leaves per tree.
- Larger values increase model complexity, while smaller values help control overfitting.

min_child_samples = 48

- Minimum number of data points needed in a leaf.
- Increasing this value can prevent the model from learning noise in the data.

subsample = 0.75

- Proportion of data to randomly sample for each boosting iteration.
- Helps prevent overfitting by adding randomness.

colsample_bytree = 0.55

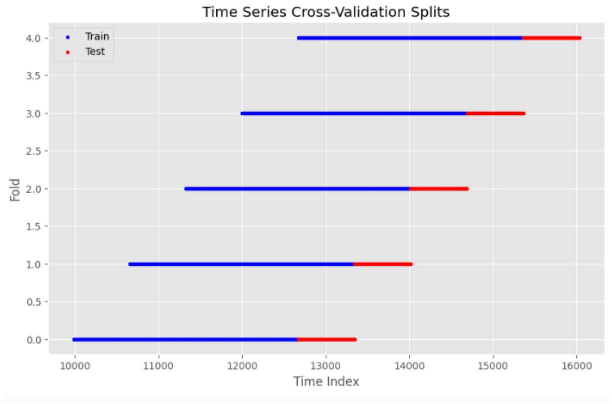
- Fraction of features to use for each tree.
- Reduces overfitting by limiting the number of features each tree can use.

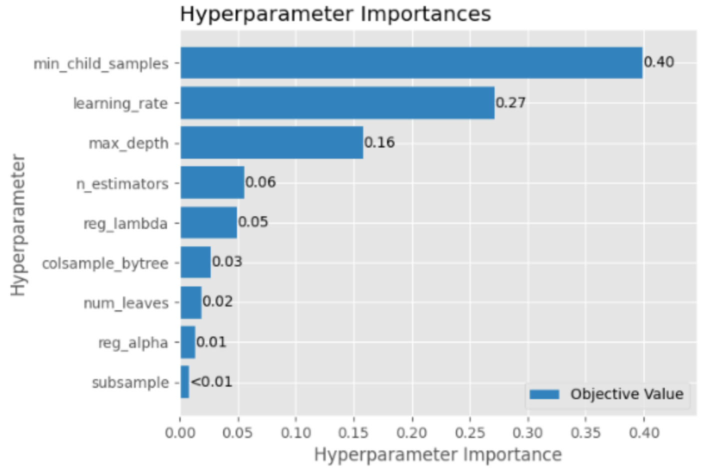
reg_alpha = 0.45

- L1 regularisation term on weights.
- Helps reduce model complexity and prevent overfitting by adding a penalty for large coefficients.

reg_lambda = 0.4

- L2 regularisation term on weights.
- Controls model complexity by penalising large weights, which helps prevent overfitting.

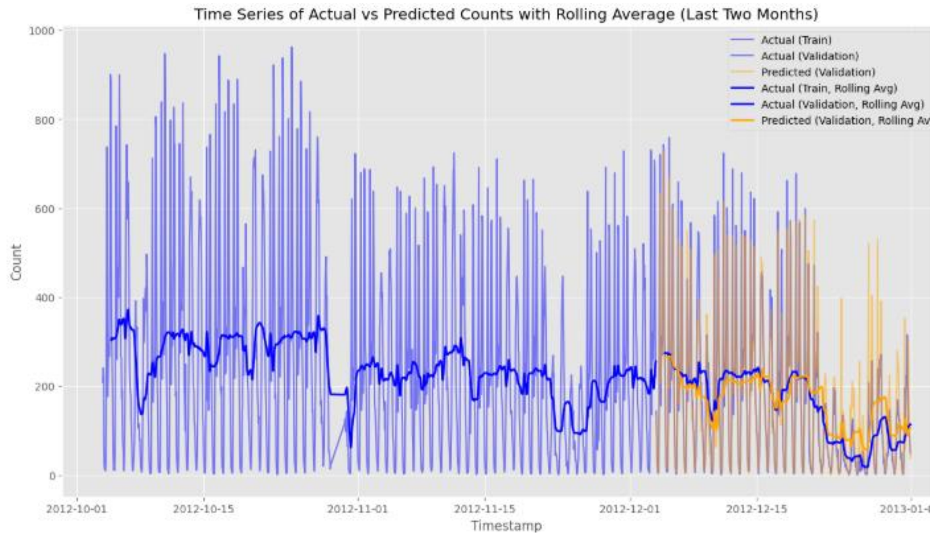




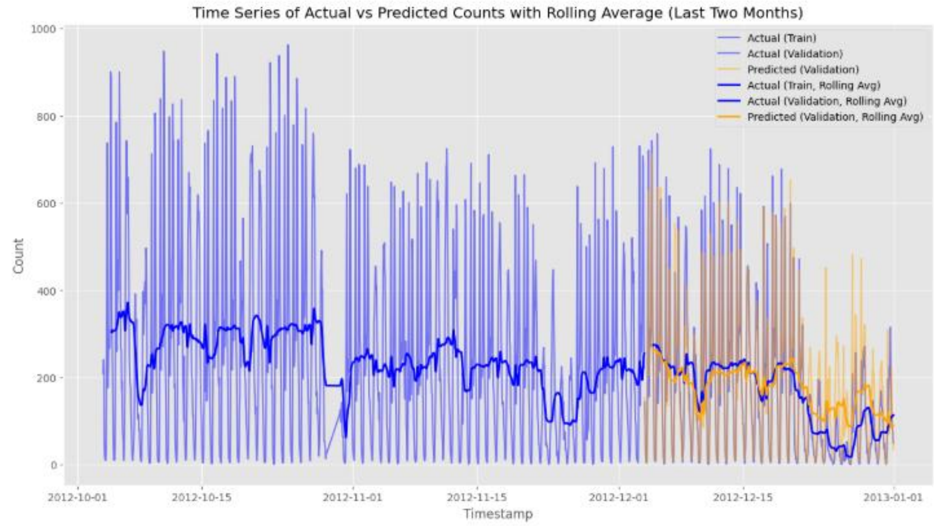
Final Model Results

- The final RMSE was 59.2
- This improved on a baseline model with original features RMSE of 65.3
- The tuned model judges the tough Christmas period better than the basic model.

Tuned Model RMSE of 59.2



Base Model RMSE of 65.3



Next Steps

Further Analysis of User Types:

- Deepen the investigation into the behaviour of registered vs. casual users, focusing on variations across time (weekdays vs. weekends).
- Explore additional factors that could influence casual user behaviour, such as specific events, local festivals, or school holidays.

Refine and Add Features:

- Incorporate features such as **rental price**, **payday**, **alternative transport costs**, **fuel prices**, and **congestion charges** (e.g., ULEZ enforcement).
- Investigate **location-specific demand**, allowing for targeted planning of bike distribution and capacity management.

Quality and Completeness of Data:

- **Examine holidays data** further to ensure accuracy or potentially remove it if it's unreliable.
- Assess **weather data quality** (e.g., missing or incorrect data) to improve model robustness.

Model Improvement and Monitoring:

- Continue **model tuning** and **validation** by testing additional machine learning models and advanced techniques like **ensemble models** and combining SARIMA models.
- Develop **predictive models for different time frames** (e.g., next 4 weeks, seasonal predictions) and assess long-term performance.

Explore Business Impact:

- **Evaluate the effect of targeted promotions** on demand, especially for casual users.
- Consider implementing a **location-based bike distribution system** to maximise revenue opportunities and reduce missed rentals.