Project Report: Predict Bike Sharing Demand with AutoGluon

# 1. Introduction

This project aims to predict the hourly demand for bike rentals using the dataset from the Kaggle "Bike Sharing Demand" competition. The problem involves structured (tabular) data and requires predicting a continuous target (`count`) using historical demand and contextual features such as time, weather, and temperature. We use AutoGluon, an AutoML framework, to automate model selection, feature engineering, and hyperparameter tuning.

# 2. Data Loading and Exploration

a. Downloading and Loading Data:  
- The dataset was downloaded from Kaggle using the Kaggle CLI and API token.  
- All datasets (train.csv, test.csv, sampleSubmission.csv) were loaded into pandas DataFrames.

# 3. Feature Creation and Data Analysis

## a. Feature Engineering

- Extracted `hour` from `datetime`.  
- Added as a feature to both train and test datasets.

## b. Exploratory Data Analysis (EDA)

- Plotted histograms for all features using matplotlib.  
- Identified distributions and potential anomalies.

## c. Data Type Conversion

- Converted `season` and `weather` columns to categorical types.

# 4. Model Training With AutoGluon

## a. Initial Model Training

- Used TabularPredictor with label `count`.  
- Ignored columns: `casual`, `registered`.  
- Evaluation metric: RMSE.  
- Presets: best\_quality.  
- Time limit: 600 seconds.

## b. Model Evaluation

- Used `fit\_summary()` and leaderboard for evaluation.

# 5. Improved Model with Feature Engineering

- Retrained model after feature engineering.  
- Improved validation and Kaggle scores.

# 6. Hyperparameter Optimization

- Tuned XGBoost, CatBoost, LightGBM using AutoGluon's `hyperparameters` and `hyperparameter\_tune\_kwargs`.  
- Used random search across defined hyperparameter spaces.

# 7. Model Predictions and Kaggle Submission

- Generated predictions on the test set.  
- Negative values were set to zero.  
- Submission prepared and sent via Kaggle CLI.

# 8. Results and Model Comparison

a. Model Performance Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Version | Features Added | Hyperparameters Tuned | Validation Score (RMSE) | Kaggle Score |
| Initial (baseline) | None | Default | 52.8 | 1.81 |
| + Hour Feature | hour | Default | 31.1 | 0.75 |
| + HPO | hour | XGB, CAT, GBM | 39.5 | 0.49 |

b. Hyperparameter Table

|  |  |  |
| --- | --- | --- |
| Model | Hyperparameters Tuned | Kaggle Score |
| Initial | Default | 1.81 |
| Features | Default | 0.75 |
| HPO | XGB, CAT, GBM | 0.49 |

# 9. Discussion

- Feature Engineering Impact: Adding the `hour` feature improved score significantly.  
  
- EDA Impact: Histograms helped verify data quality.  
  
- Hyperparameter Tuning Impact: Tuning improved generalization and reduced RMSE.  
  
- Best Model: Identified via AutoGluon leaderboard and validation scores.

# 10. Conclusion

AutoML with AutoGluon streamlined tabular modeling. Feature engineering and hyperparameter tuning contributed to leaderboard improvements. Future work could explore external data sources or time-series-aware modeling.

# 11. References

- Kaggle Bike Sharing Demand Competition: https://www.kaggle.com/c/bike-sharing-demand

- AutoGluon Documentation: https://auto.gluon.ai/