The University of Texas at Austin
Cockrell School of Engineering

NOVEMBER 30, 2023

PREDICTING DEMAND FOR BIKESHARING IN WASHINGTON DC

Semester Presentation for ECE 381.3: Applied Machine Learning

LAKSHYA JAGADISH, FADEEL KHAN, ASVIN KUMAR, POWELL LOWE
The University of Texas at Austin

BACKGROUND: CAPITAL BIKESHARE

- Capital Bikeshare: government-run bike share company with **3.4 million annual bike ridership** (~9k daily riders) in Washington D.C. and surrounding areas
- Riders can buy annual memberships (**"registered"** riders) or pay per trip (**"casual"** riders)
- Understanding variations in ridership an important consideration for future government city development
  - City planning
  - Strategies for advertisements on bikes
  - Strategies for marketing for more ridership

Figure Adapted from Capital Bikeshare

DATA DESCRIPTION

DATASET BREAKDOWN

- Time series data, aggregated by both day and hour
- Feature set:
  - Time series (date, day, year, month, hour)
  - Seasonal information
  - Weather information (humidity, windspeed, temperature, state of weather)
  - Day information (weekday, weekend, holiday)
- Target variables:
  - “Registered” riders using bikes
  - “Casual” riders using bikes
  - Total riders (registered + casual)

- **day.csv** (aggregated by day between 2011-2012)
  - Size: 731 samples x 16 columns
- **hour.csv** (aggregated by hour between 2011-2012)
  - Size: 17379 samples x 17 columns

**all features same for both except for “hour” feature in hour.csv*

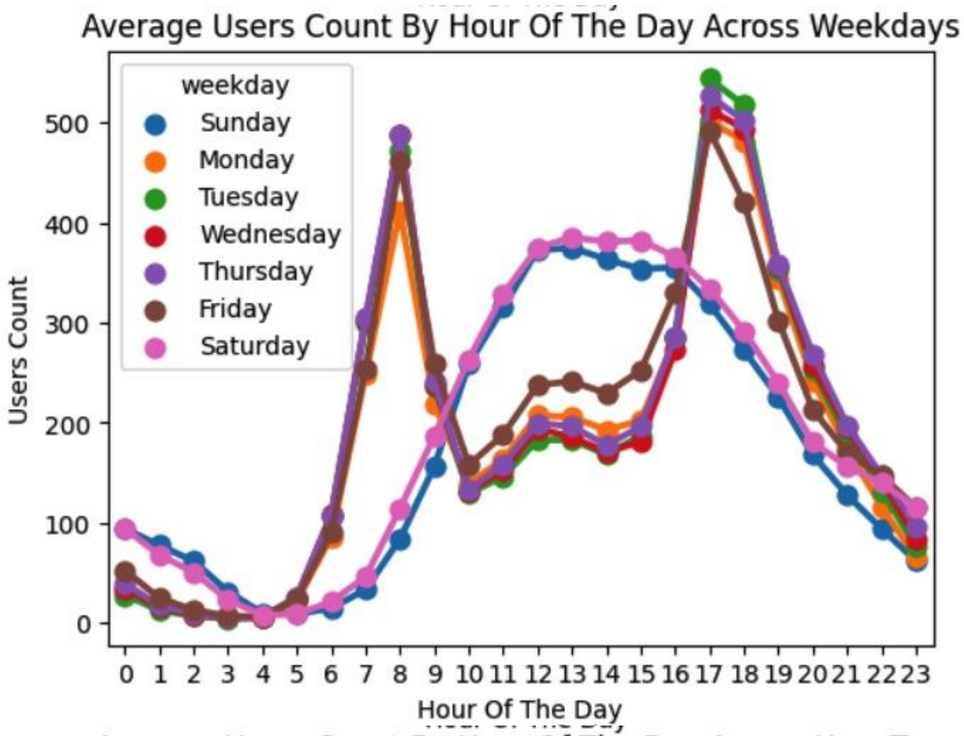
PREVIOUS WORK

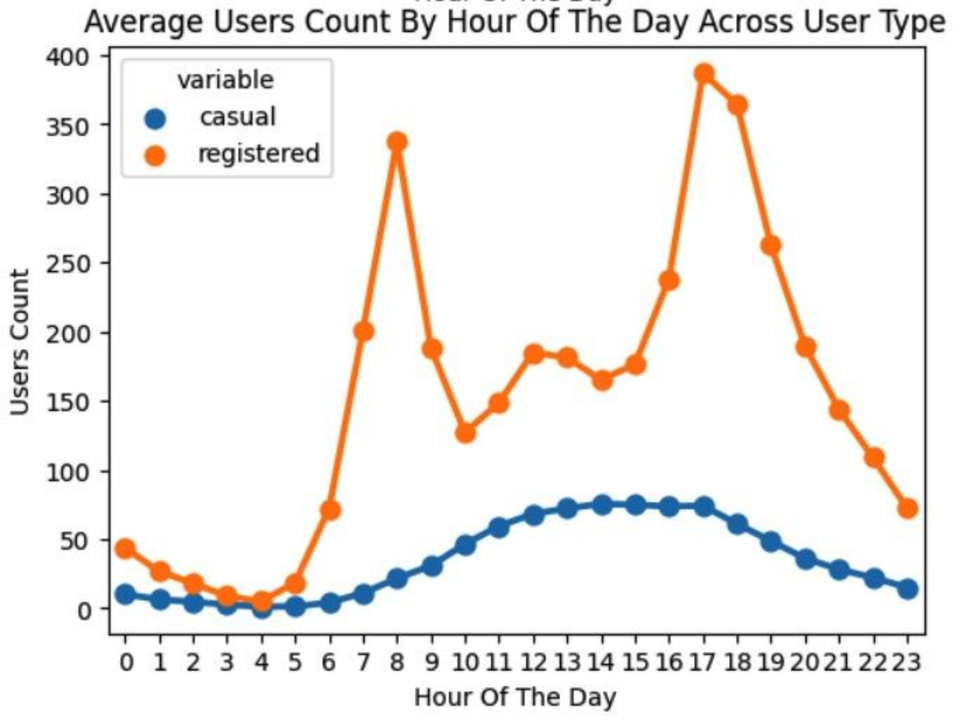
- Recurrent Neural Network (RNN) regression for predicting rider counts (Petnehazi et al.)
  - R^2 = 0.78-0.96
- Decision Trees multi-class prediction of seasons based on rider count and other features (Al-Otaibi et al.)
  - Classification accuracy = 0.4-0.7
- Kaggle Competition
  - Non-neural network models (LR, ensemble methods); R^2 = ~1.0
  - No model/feature interpretation
- Statistical visual interactions between features (Britton et al.)
  - No predictive modeling

EXPLORATORY DATA ANALYSIS

- Multiple seasonalities
  - Weekends vs Weekdays
  - Winter vs Summer
  - Morning/Evening versus Midday
- Commutes dominated by Registered Users
- Temperature is most correlated weather feature with Count

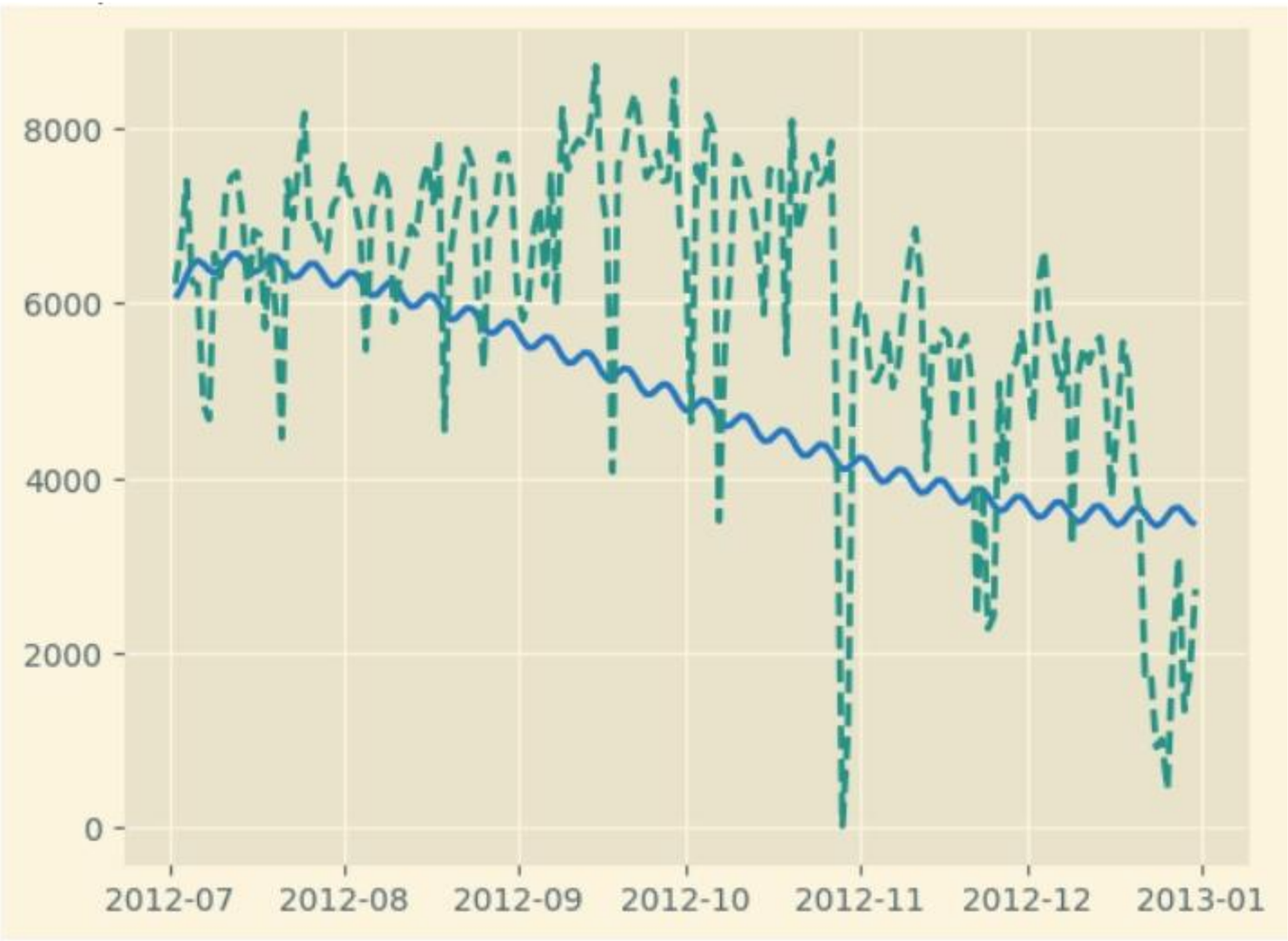
  Humidity and Wind Speed → minimal correlation





TBATS Baseline

- Incorporates:
  - Trigonometric seasonality
  - Box-Cox Transformation
  - ARMA
  - Trend
- Pros:
  - Multiple seasonalities
  - Performs exponential smoothing
- Poor performance and extremely slow



PRE-PROCESSING

- Normalization of feature set
- Seeding data, 80-20 train-test split
- 3-fold cross validation
- GridSearchCV for hyperparameter tuning
- Top 3 principal components explained 94-97% of variance → data inverse-transformed based on PCA
  - Model evaluation pre- and post-PCA

MODEL SELECTION

- Linear Regression
- Support Vector Regression
- Random Forest
- XGBoost Forest
- Recurrent Neural Network (Long Short-Term Memory)

Emphasis placed on the simplicity and interpretability of models

MODEL RESULTS – REGISTERED RIDERS

- hourly dataset

Model	RMSE	R^2
Linear Reg.	831.3	0.83
Random Forest Reg	10.64	0.99
RF Reg (w/ PCA)	10.14	1.00
XGBoost	36	0.95
XGBoost (w/ PCA)	1.90	1.00
LSTM	117.053	0.613
LSTM (w/ PCA)	112.747	0.641

- daily dataset

Model	RMSE	R^2
Linear Reg.	925.06	0.76
Random Forest Reg	80.11	1.00
RF Reg (w/ PCA)	81.02	1.00
XGBoost	523	0.90
XGBoost (w/ PCA)	42.46	1.00
LSTM	915.016	0.664
LSTM (w/ PCA)	968.066	0.624

MODEL RESULTS – CASUAL RIDERS

- hourly dataset

Model	RMSE	R^2
Linear Reg.	36	0.88
Random Forest Reg	4.89	0.99
RF Reg (w/ PCA)	4.91	0.99
XGBoost	14.5	0.91
XGBoost (w/ PCA)	1.08	1.00
LSTM	20.129	0.871
LSTM (w/ PCA)	20.138	0.871

- daily dataset

Model	RMSE	R^2
Linear Reg.	342	0.71
Random Forest Reg	144.57	0.97
RF Reg (w/ PCA)	116.33	0.97
XGBoost	251	0.84
XGBoost (w/ PCA)	38.92	1.00
LSTM	450.143	0.591
LSTM (w/ PCA)	483.997	0.527

KEY TAKEAWAYS + INTERPRETATION

- SHAP is a game theoretic approach to explain the interpretability of any machine learning model.
- SHAP values describe how each feature of each input modifies a “base rate” to get the target value.
- Generally, higher absolute values have higher impac on model output.

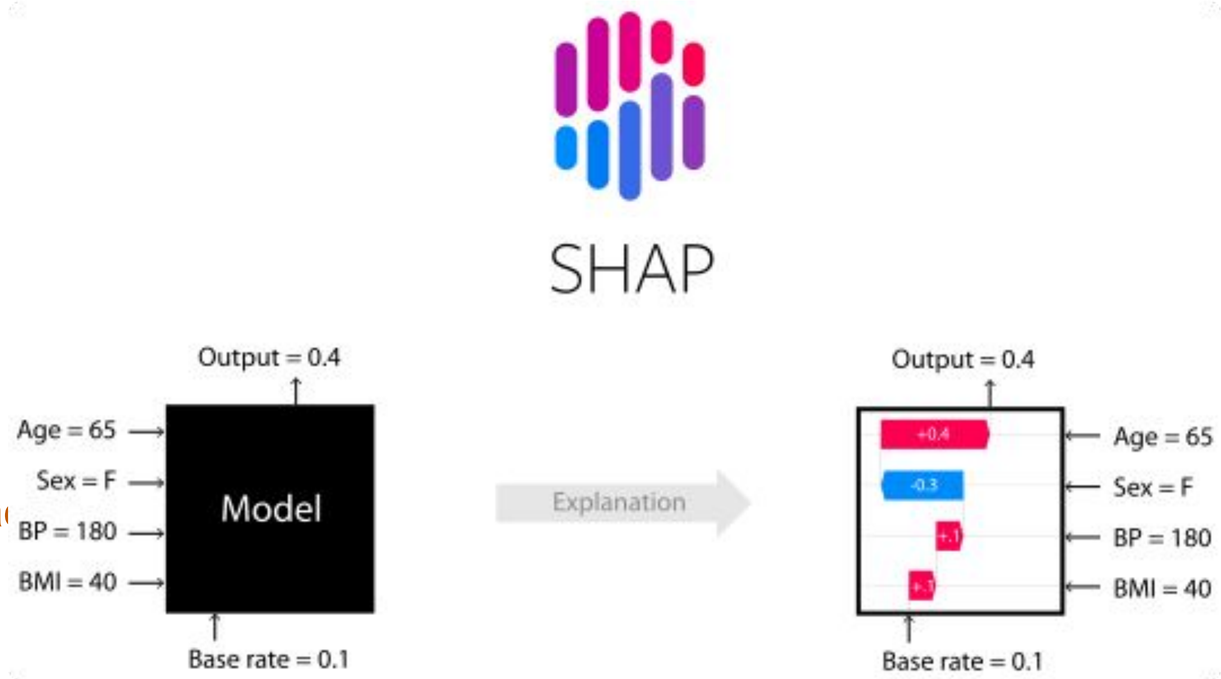
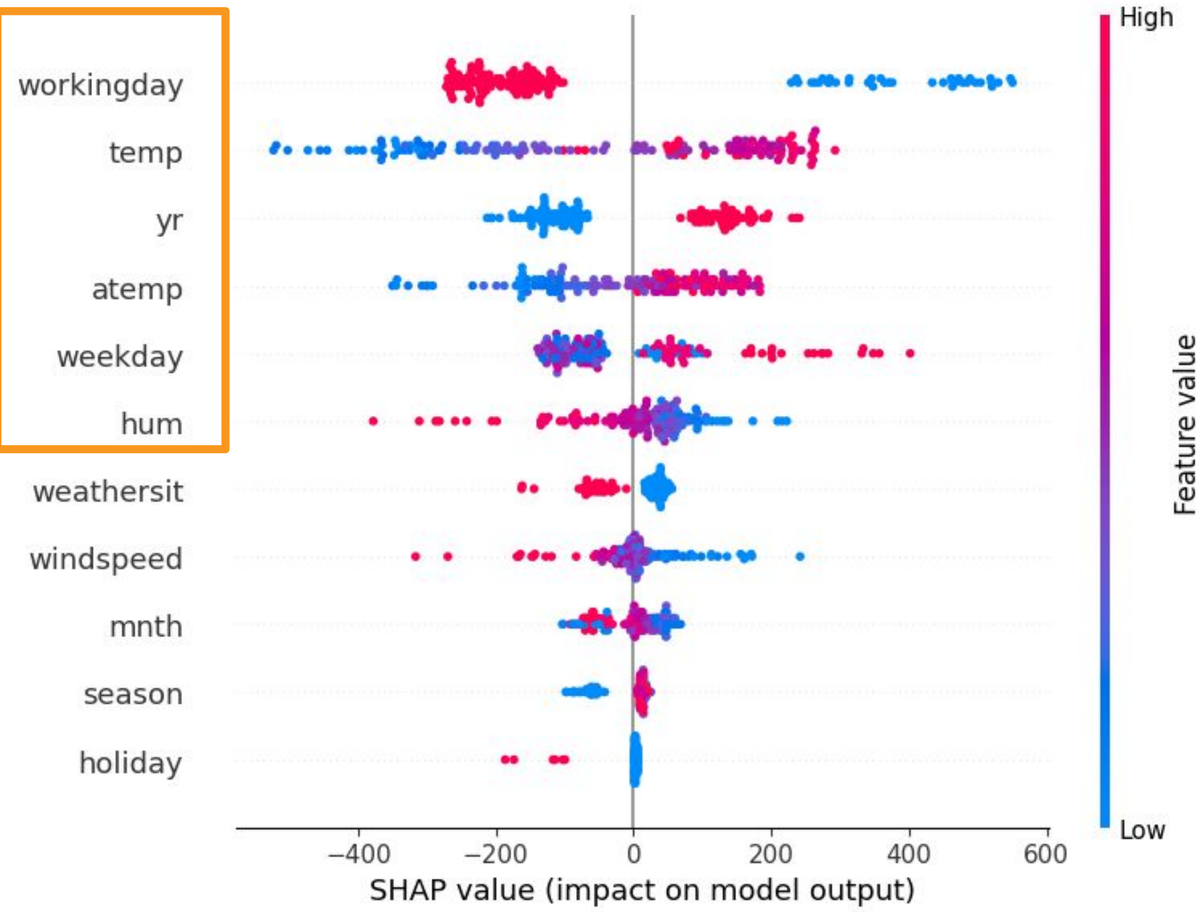


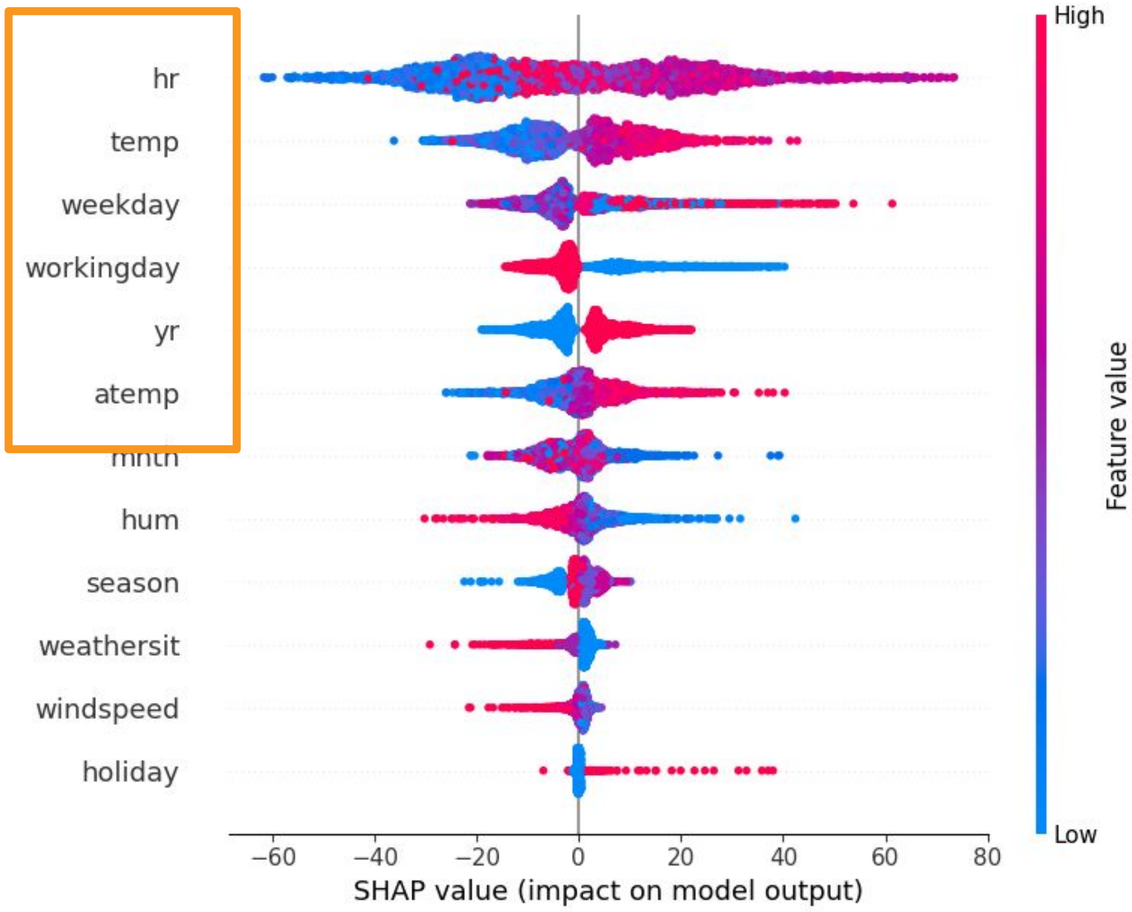
Image from SHAP Repository

XGBOOST - CASUAL

Day

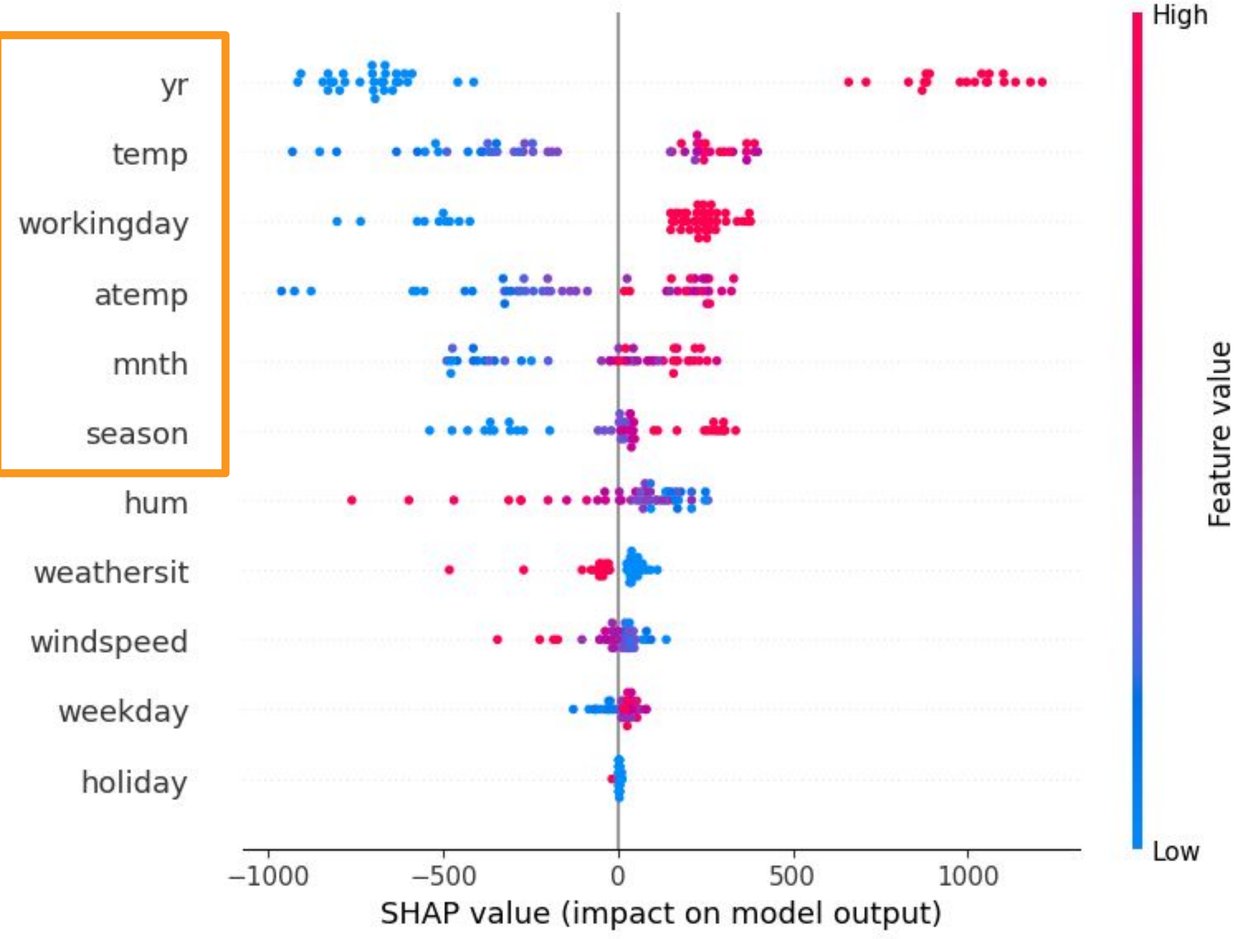


Hour

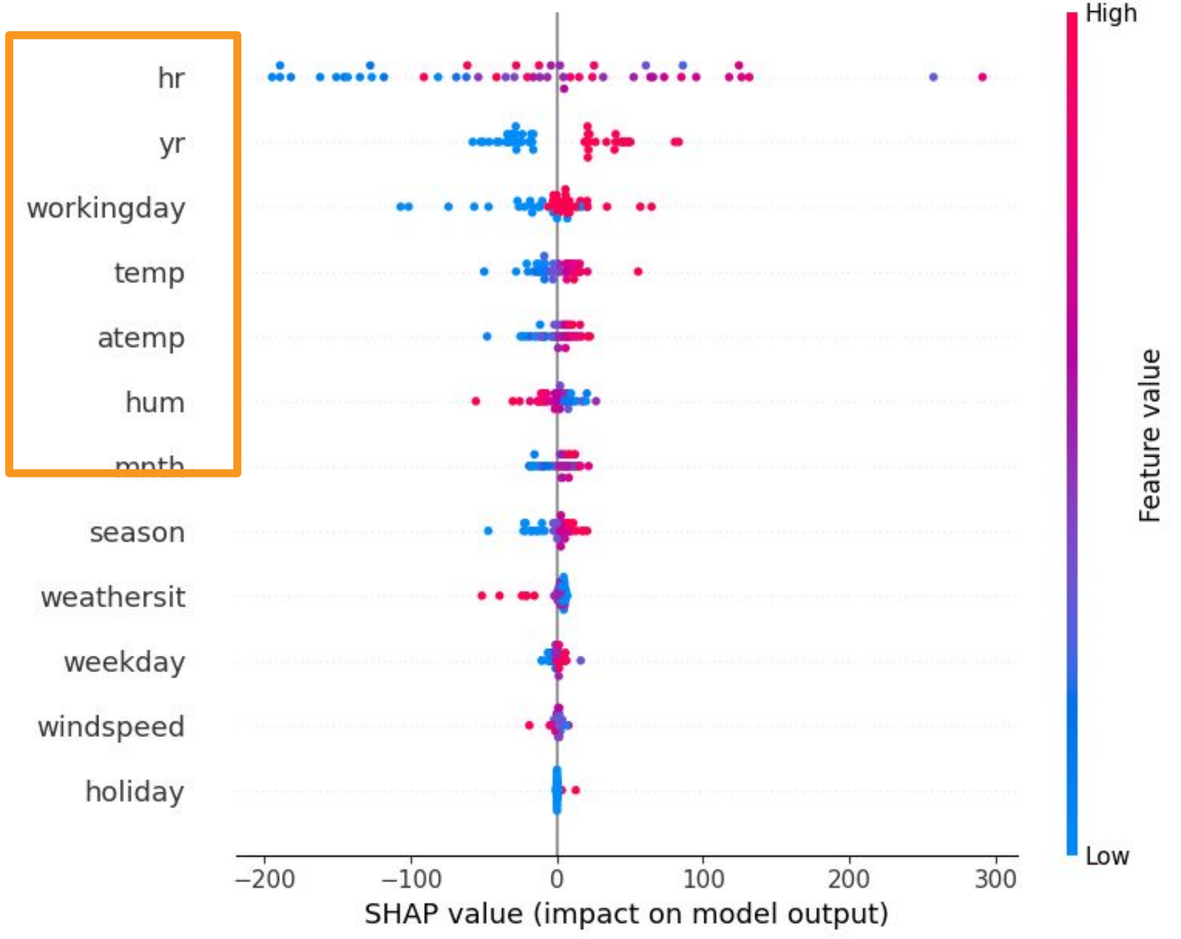


RANDOM FOREST - REGISTERED

Day

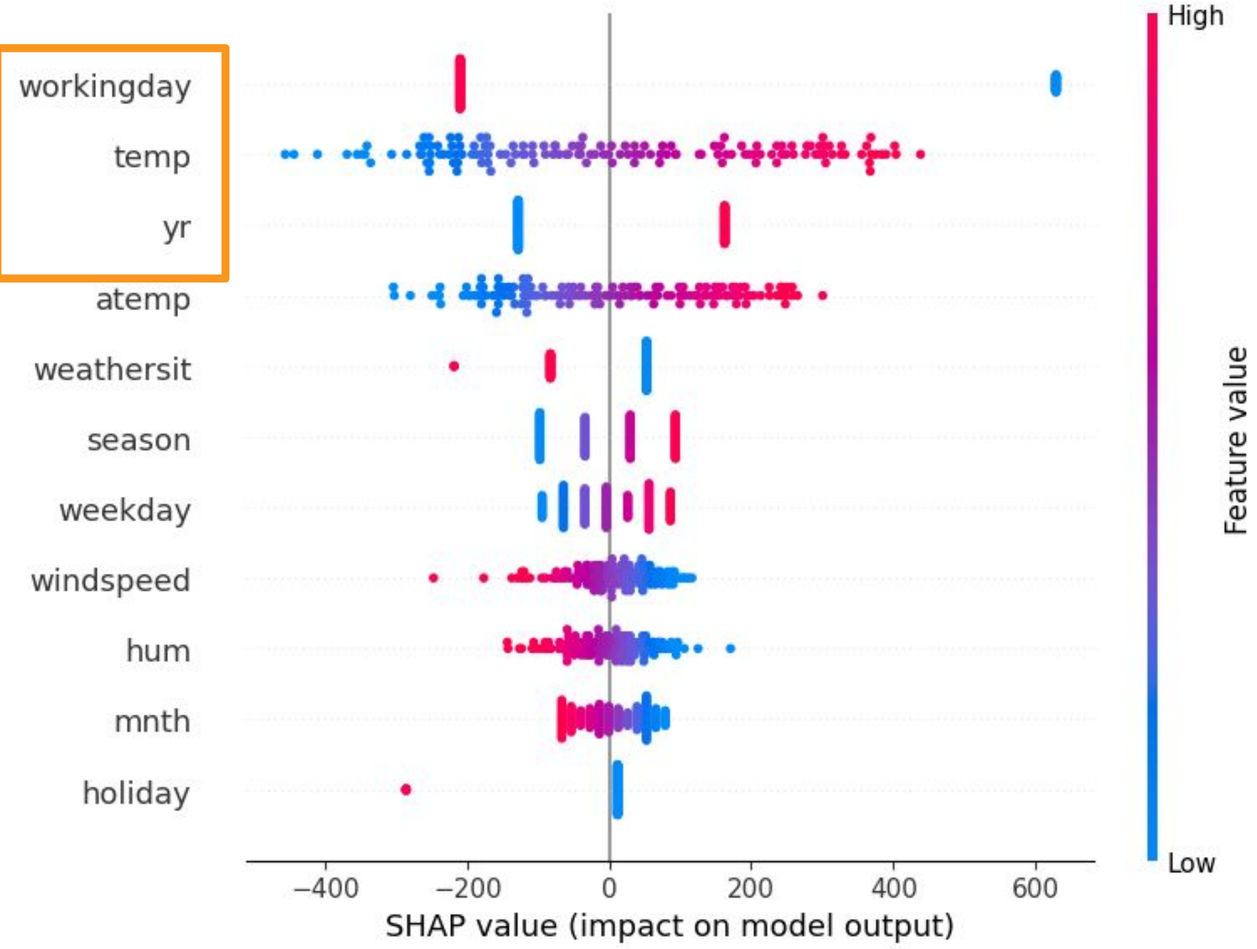


Hour

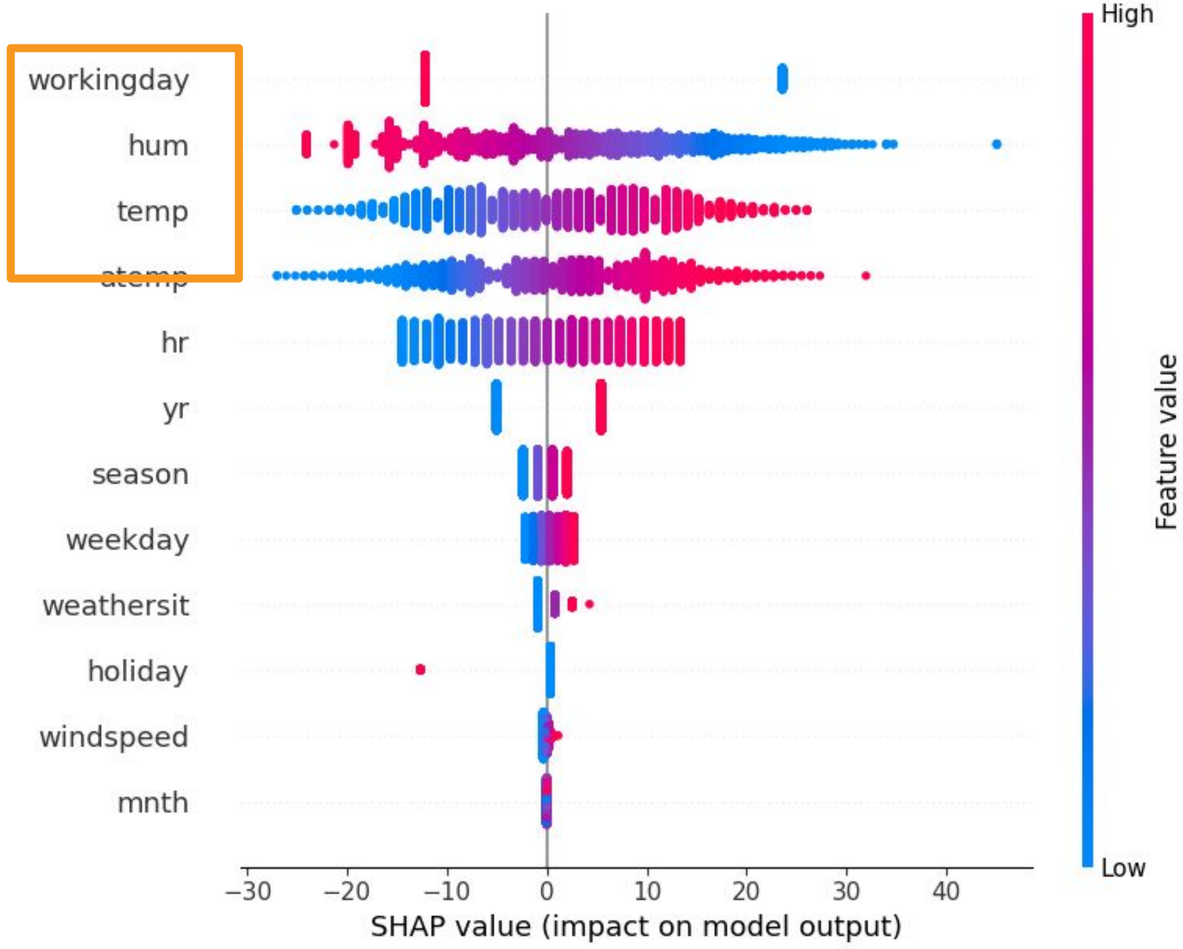


LINEAR REGRESSION - CASUAL

Day

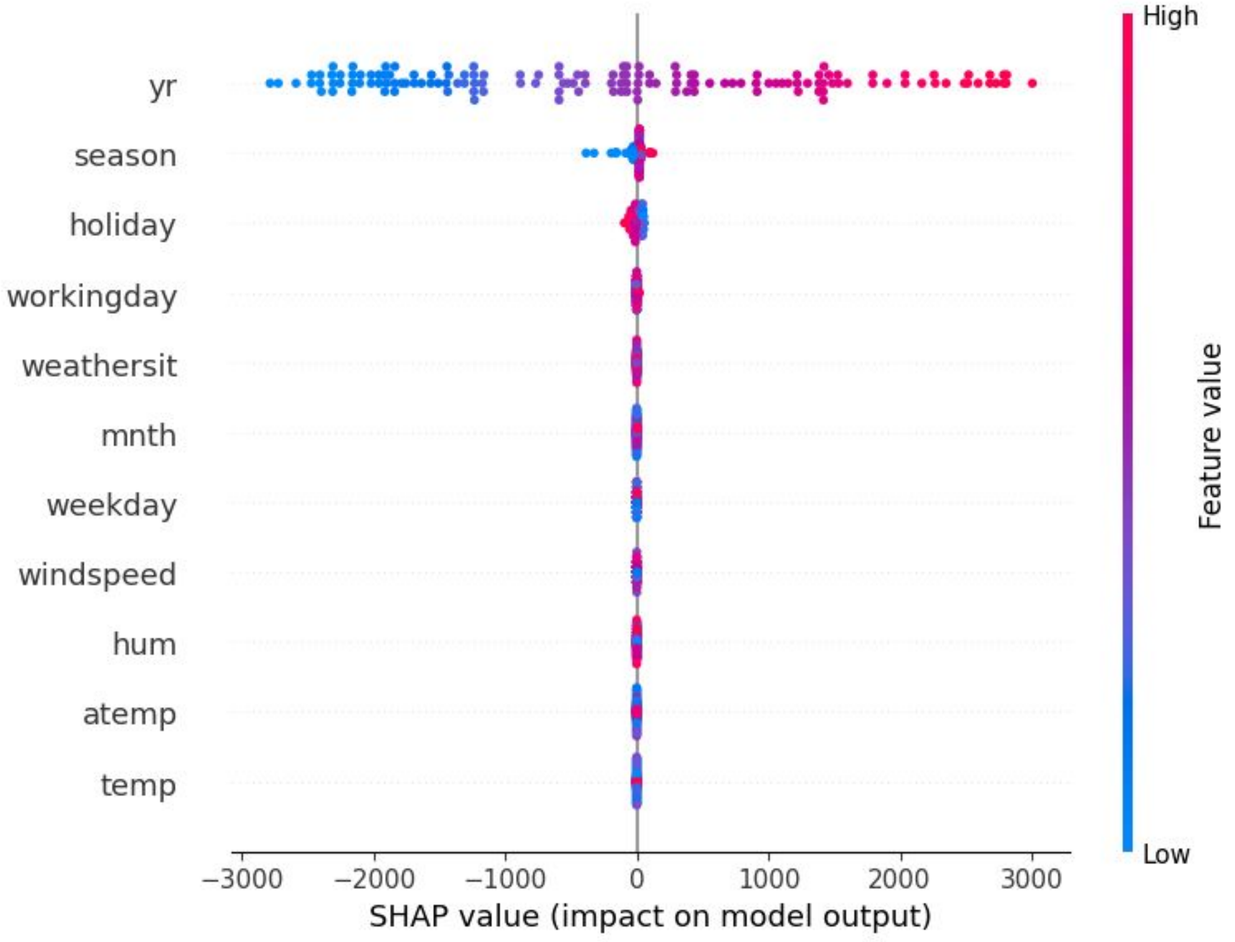


Hour

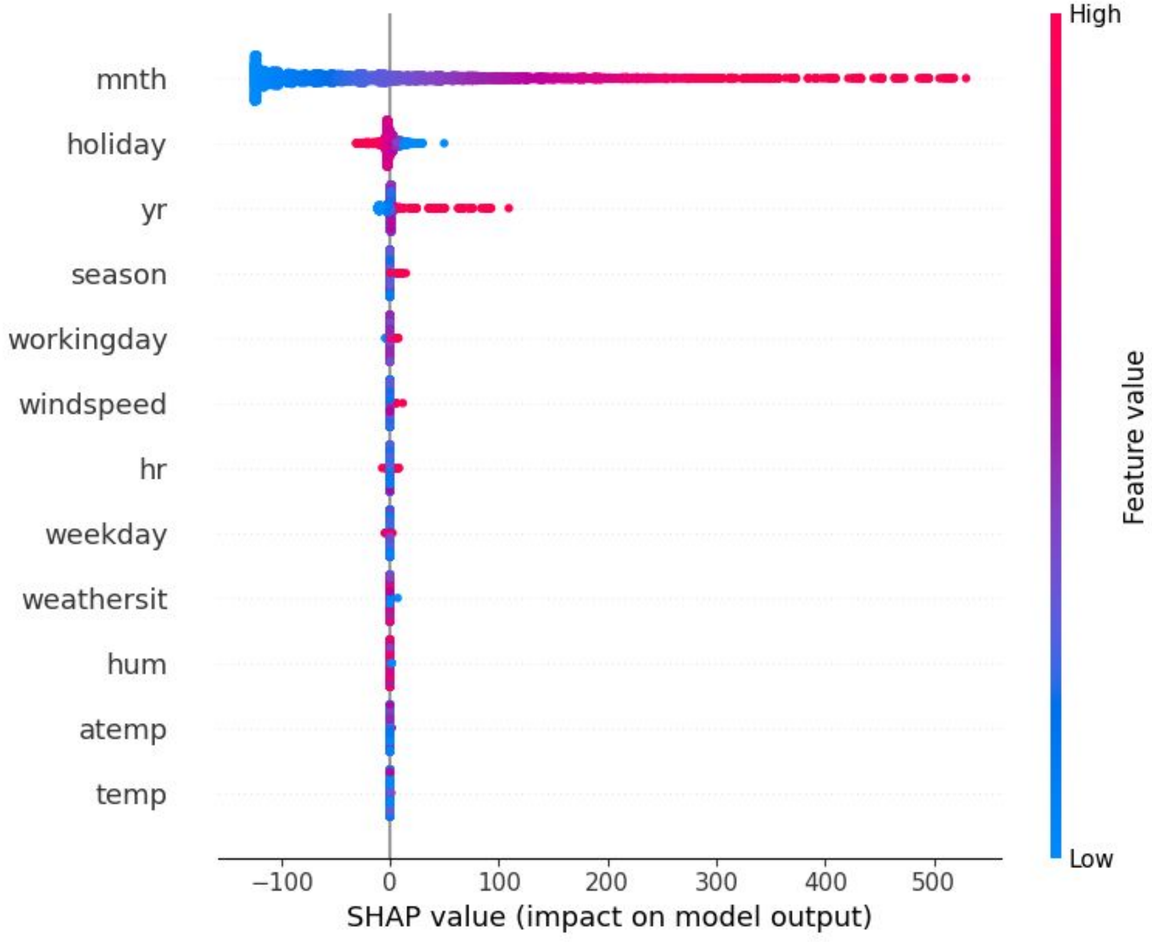


XGBOOST (PCA) - REGISTERED

Day



Hour



KEY LIMITATIONS

DATASET

- Size (2 years) limits training of yearly trend behavior

MODELS

- TBATS is difficult to train and interpret
- Hard to tell whether these models will perform on longer-duration data

FUTURE WORK

- Continuously validating and updating the model as new data becomes available - the current model is dated at 2011, it is important to keep up with the current trends.
- Real-time prediction capabilities - learn to adapt to sudden changes in demand.
- Explore use of RL for optimizing bike-sharing operations over-time.
- Integrate other data sources such as social media or locations information which can provide insights into user behaviour, sentiments and preferences.
- Using ensemble models to combine the inferences from different models could be interesting to see how the model can perform in comparison to the XGBoost model.
- Model interpretation - Could explore other frameworks like LIME and ELI5.

REFERENCES

- Lichman, M.: UCI machine learning repository (2013). http://archive.ics.uci.edu/ml
- M. Britton, ‘VINE: Visualizing statistical interactions in black box models’, *arXiv [cs.LG]*, 01-Apr-2019.
- A. Appice et al. (Eds.): ECML PKDD 2015, Part I, LNAI 9284, pp. 184–199, 2015. DOI: 10.1007/978-3-319-23528-8 12
- G. Petneházi, ‘Recurrent neural networks for time series forecasting’, arXiv [cs.LG], 31-Dec-2018.

The University of Texas at Austin
Cockrell School of Engineering