Bike Sharing System

MinJae Song
Stephen Acurso
CIS 4567
Dr. Salehan

Introduction

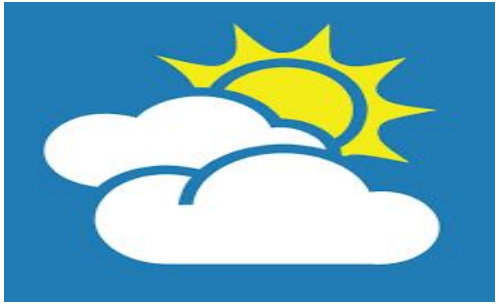
- Bikeshare Program - casual and registered users rent from a particular location and return at another location
- Daily usage analysis has been done, but we want to know hourly usage
- Objective: Predict the hourly demand of bike rentals to accordingly prepare for supply
- Want to see if there is a relationship between temperature, humidity, month, hour, and bike rentals
- Build a machine learning model to help us predict future bike rentals by analyzing variables that can possibly impact one's decision to rent and continue to improve

Data collection

- We acquired the dataset from Kaggle.com, bike sharing company based in Washington D.C., Capital bike share and data was from 2011-2012 collected hourly
- Our dataset included weather information such as temperature, “feels-like” temperature, humidity, and wind speed. Our dataset also included seasonal information consisting of hour, day of the week, and if the day fell on a holiday or not
    - Other weather information included: clear skies, cloudy, rain, snow
- Hourly rentals included registered/casual users
    - These were aggregated to the column “count” (dependent variable)
- Our dataset did not include any missing or null values
    - We did have outliers however we decided to include them in our model
- We did not eliminate any data points from our dataset

Descriptive analysis

- Total of 17379 records spanning 17 columns
- Average rentals per hour: 189.46
- The hours with highest average rentals included 8am, 5pm, 6pm (Times most people either begin or end workday)
- We also noticed a relationship between weather and number of rentals

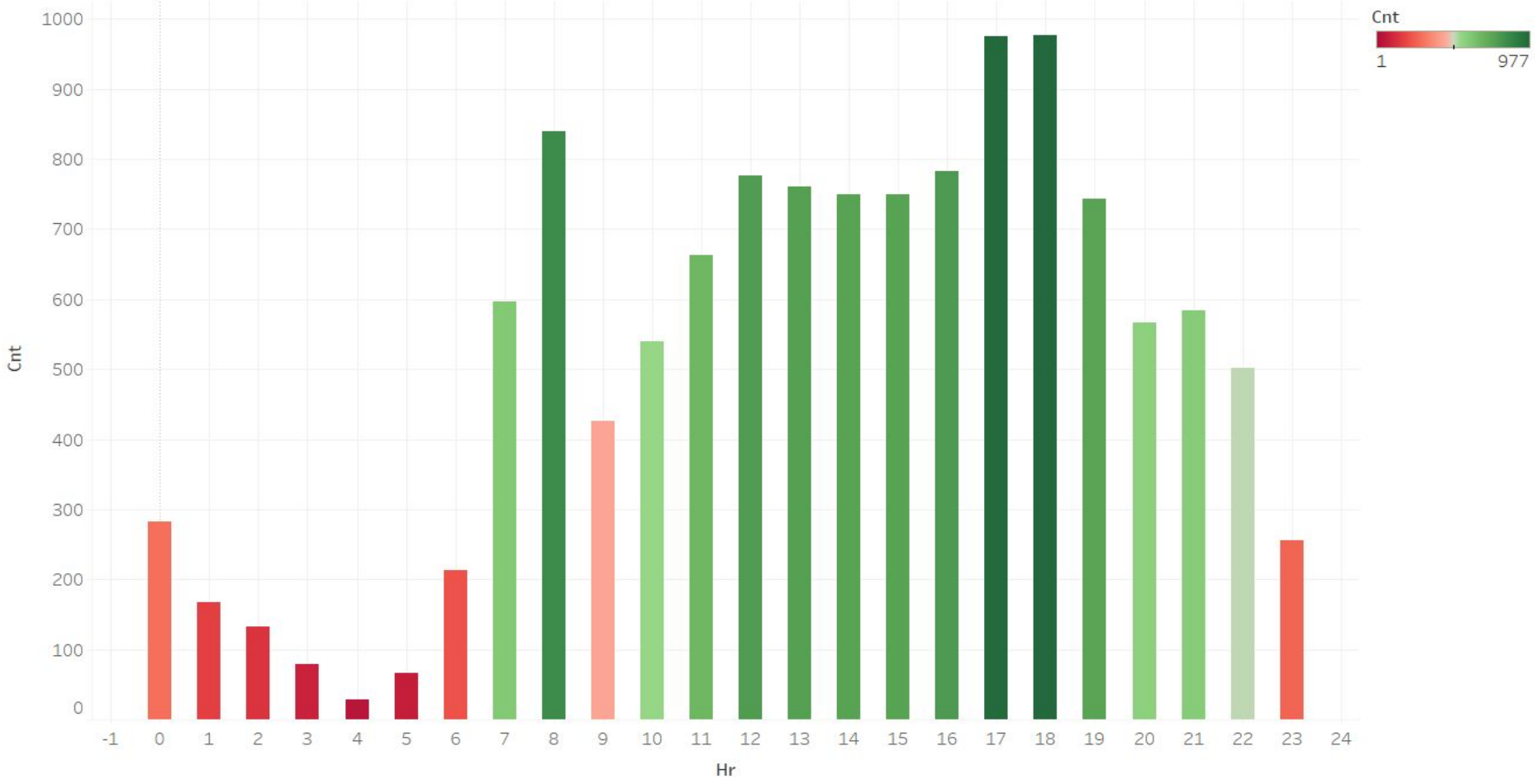


Descriptive analysis

instant	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
1	2011-01-01	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0	3	13	16
2	2011-01-01	1	0	1	1	0	6	0	1	0.22	0.2727	0.8	0	8	32	40
3	2011-01-01	1	0	1	2	0	6	0	1	0.22	0.2727	0.8	0	5	27	32
4	2011-01-01	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0	3	10	13
5	2011-01-01	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0	0	1	1
6	2011-01-01	1	0	1	5	0	6	0	2	0.24	0.2576	0.75	0.0896	0	1	1
7	2011-01-01	1	0	1	6	0	6	0	1	0.22	0.2727	0.8	0	2	0	2
8	2011-01-01	1	0	1	7	0	6	0	1	0.2	0.2576	0.86	0	1	2	3
9	2011-01-01	1	0	1	8	0	6	0	1	0.24	0.2879	0.75	0	1	7	8
10	2011-01-01	1	0	1	9	0	6	0	1	0.32	0.3485	0.76	0	8	6	14

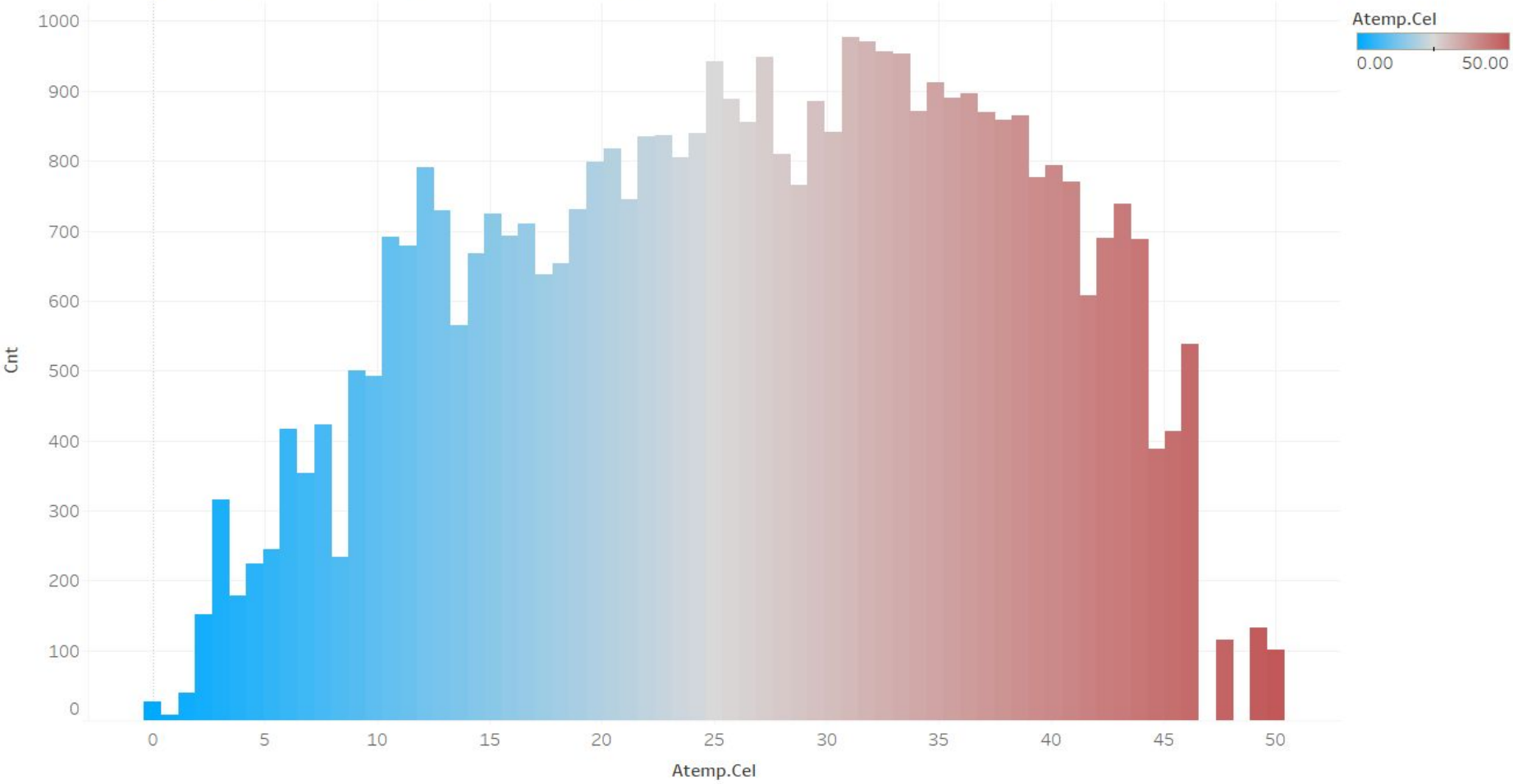
Our first record was on 2011-01-01 00:00, and there were 16 bikes rented!

Relationship between hr and bike rentals



Average count related to month

Relationship between air temp and bike rentals



Data Preparation and Feature Engineering

- No missing values in dataset, and weather-related variables were normalized
- Data consisted of numerical and some categorical variables
- Numeric variables: temperature, feels-like temperature, humidity, windspeed, and count
- Categorical variables: season, month, hour, holiday, workingday, and weather situation
- Variables used: feels-like temperature, humidity, windspeed, count, month, hour, workingday, and weather situation
- Temperature was not needed as feels-like temperature had similar values, correlations. Holiday and day of the week was not needed.
- We used a VectorIndexer for our categorical variable and a VectorAssembler to combine our features together.
- 75/25 random split
- Ready for training!

Data Modeling

- Linear Regression is a good baseline model for prediction when dealing with a dependent variable and independent variables
- After training the model and testing it with our test data, we evaluated our model by comparing it with the actual count and the predictions. Our model had a r2 score of 0.34, meaning that it explains 34% of the variability of the data, and with a RMSE (root mean squared error) of 144.93, which is the average of the difference between the prediction and actual observation
- Not the best model, score, but we can improve. ** add error rate
- To improve, we implemented the Decision Tree regression model because this model uses branches (nodes) for prediction.
- After training, testing, and evaluating, the Decision Tree model had a r2 score of 0.597, meaning that our model explains 59.6% of the variability of the data, and with a RMSE of 113.1, average difference between prediction and observation
- We still wanted more improvement, therefore, we added a Gradient Boost to our Decision Tree model. This was a better choice than a Random Forest because it will train one tree at a time with minimal loss.
- R2 score improved to 0.76, meaning that it explains 76% of the variability of that data, and with a RMSE of 87.56, average difference between prediction and observation

Summary of Findings

- We were able to continually improve our model by implementing different models
- Linear regression was a good baseline model to improve on
- Model had made predictions of negative rentals, but we know that can't happen.
- With Decision tree regression, we were able to find the feature importance
    - Hour and Feels-like temperature had most weight
- The Gradient Boosted Tree model was the most accurate (r2=0.76, RMSE=87.56)
- Gradient Boost has many hyperparameters to work with, so we can test other settings
- For a more accurate model, we realized that it would've been better to develop two models: one to predict casual rentals and another to predict registered rentals. Adding both predictions could have been more accurate than predicting for the total rentals because registered cyclists rent more for different reasons than casual cyclists.

Implications

- By accurately predicting the number of rentals on an hourly basis we can reduce operations and maintenance cost
- Running App on AWS we can use our model to reserve resources at a much lower rate than on demand pricing, saving money in the long run
- Costs can be cut back when demand is predicted to be low by reducing the need for maintenance and reducing the number of bikes available for rent

Limitations

- Data set was collected for two year span 2011-12. Newer version and larger dataset would help predict future rentals more precisely
- More variables such as length of rental and location would improve accuracy of our prediction models