Capstone Project - 2

Bike Sharing Demand Prediction

Presented By

Rupesh Sharma

Contents

- ❑ Introduction
- ❑ About the dataset
- ❑ Data Dictionary
- ❑ Data Analysis & Visualisation
- ❑ Evaluation metric
- ❑ Feature Engineering
- ❑ Feature Selection
- ❑ Model Implementation
- ❑ Feature Importance & Final model evaluation on test set
- ❑ Challenges Faced
- ❑ Conclusion

Introduction

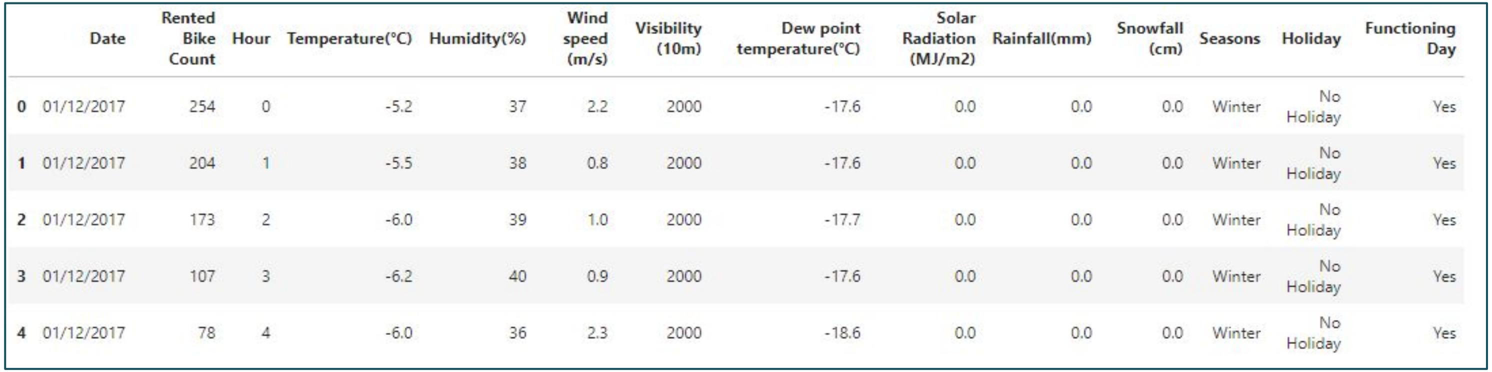
❑ A bike-sharing system is a shared transportation service in which bicycles are made available for short-term rental to individuals for a fee. This shared transportation system has grown in popularity in recent years because of its environmental benefits and ability to avoid major traffic congestion.

❑ The objective of my project is to identify the that factors influence bike demand and to build a predictive model that can assist in forecasting rental bike demand at a given time to maintain a steady supply of rental bikes at stations and hence improve the rental bike business.

About the dataset

The Bike sharing demand dataset consists of weather information and the daily bike count in Seoul, and it helps to perform data-driven exploration and acquire bits of knowledge to build a predictive model that can assist in forecasting rental bike demand at a given time to maintain a steady supply of rental bikes at stations and improve the rental bike business.

It consists of 8,760 rows and 14 columns.



	Date	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Functioning Day
0	01/12/2017	254	0	-5.2	37	2.2	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
1	01/12/2017	204	1	-5.5	38	0.8	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
2	01/12/2017	173	2	-6.0	39	1.0	2000	-17.7	0.0	0.0	0.0	Winter	No Holiday	Yes
3	01/12/2017	107	3	-6.2	40	0.9	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
4	01/12/2017	78	4	-6.0	36	2.3	2000	-18.6	0.0	0.0	0.0	Winter	No Holiday	Yes

Data Dictionary

Column	Data Type	Description
Date	Object	Date in year-month-day
Rented Bike Count	Integer	Count of bikes rented at each hour
Hour	Integer	Hour of the day
Temperature	Float	Temperature in Celsius
Humidity	Integer	Humidity in percentage
Windspeed	Float	Wind Speed in m/s
Visibility	Integer	Visibility in 10m
Dew point temperature	Float	Dew point temperature in Celsius

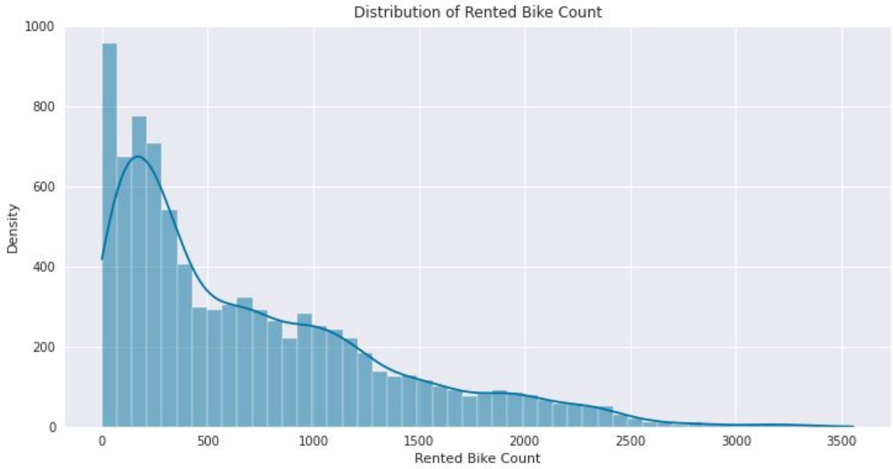
Data Dictionary

Column	Data Type	Description
Solar radiation	Float	Solar radiation in MJ/m2
Rainfall	Float	Rainfall in mm
Snowfall	Float	Snowfall in cm
Seasons	Object	Consists of four different seasons (Winter, Spring, Summer, Autumn)
Holiday	Object	Holiday or not
Functional Day	Object	Functional day or not

Data Analysis & Visualisation

- ❑ The distribution of the dependent variable, Rented Bike Count, is shown on the right.

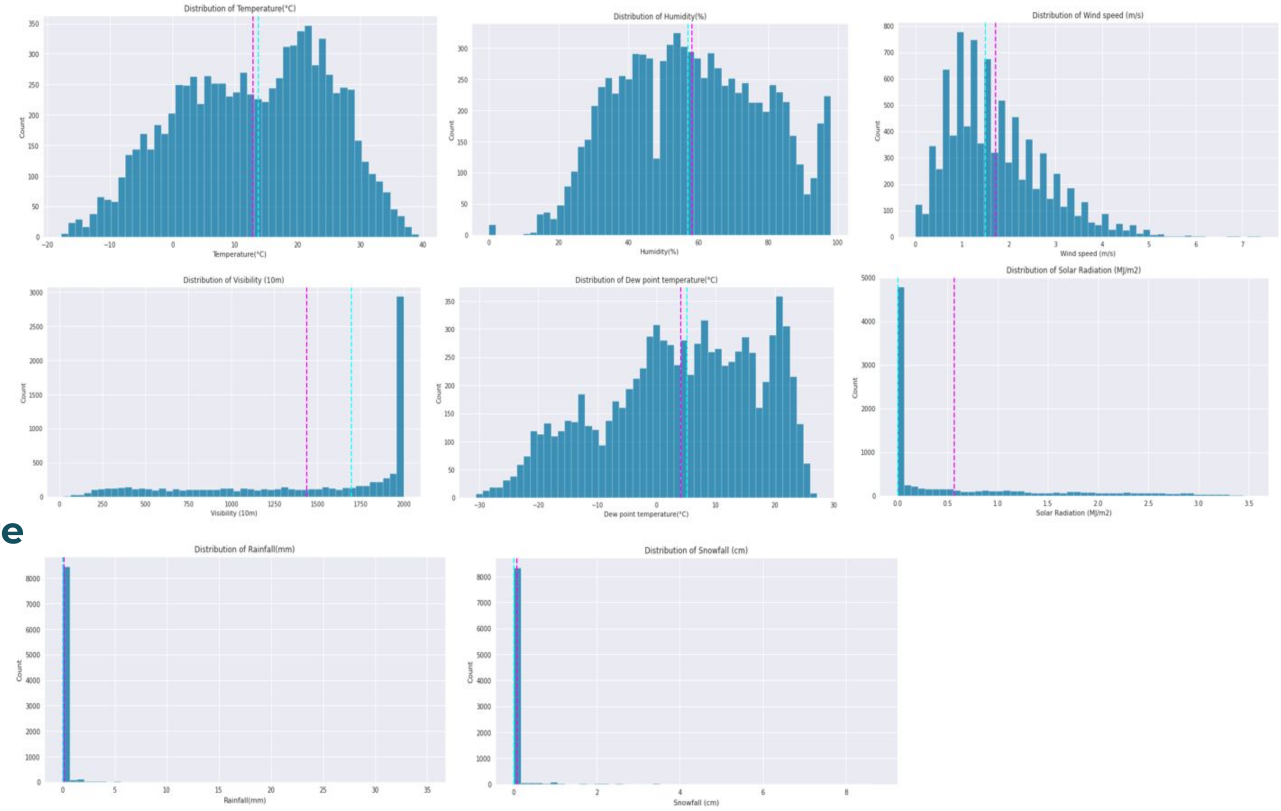
- ❑ It indicates that the dependent variable is right skewed.



Data Analysis & Visualisation

- The histogram on the right depicts the distribution of continuous features.

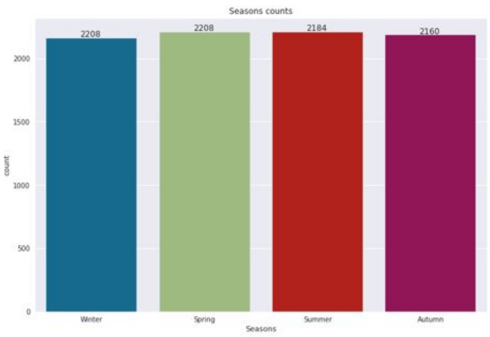
- It indicates that the continuous features are skewed.

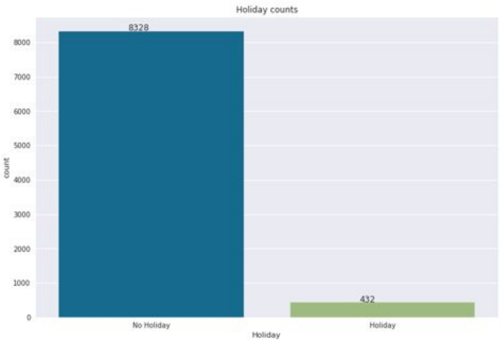


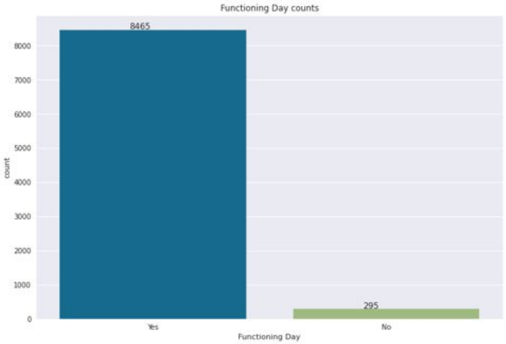
Data Analysis & Visualisation

- The countplot on the right depicts the representation of categorical features in the dataset.

- From these visualisations, we can infer that holiday and non-working days account for only 4.93 and 3.3 percent of the dataset, respectively.



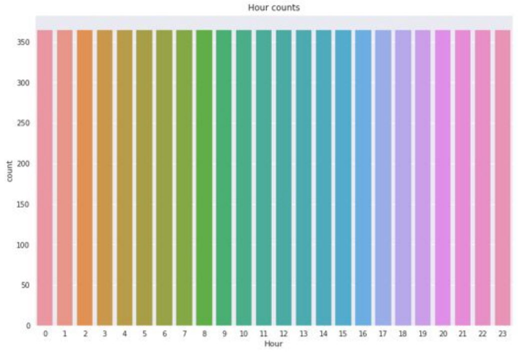


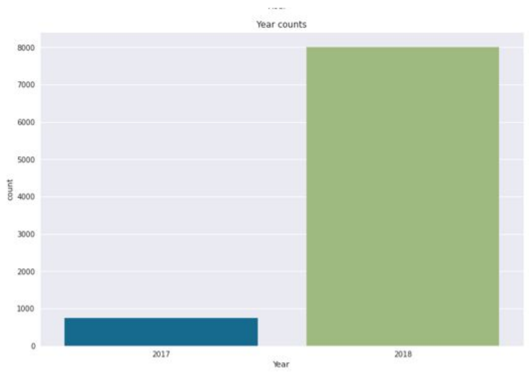


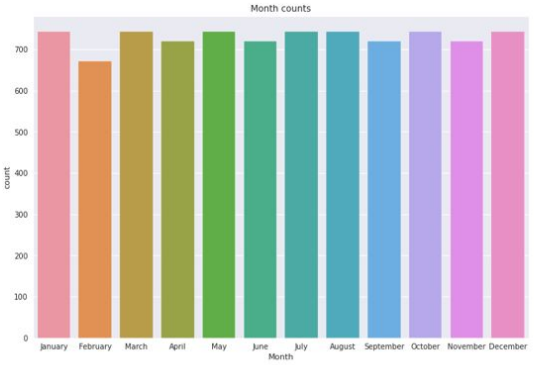
Data Analysis & Visualisation

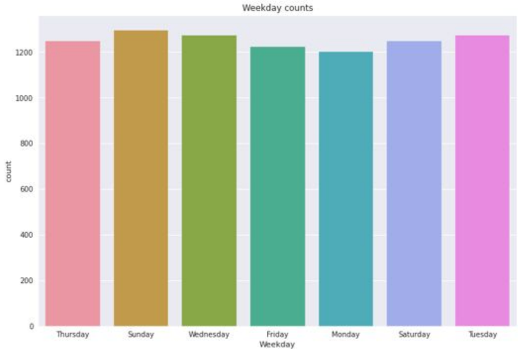
❑ The countplot on the right depicts the representation of cyclical features(time related) in the dataset.

❑ From these visualisations, we may infer that majority of data was collected in 2018 and expect for the year, all cyclical features are equally represented.







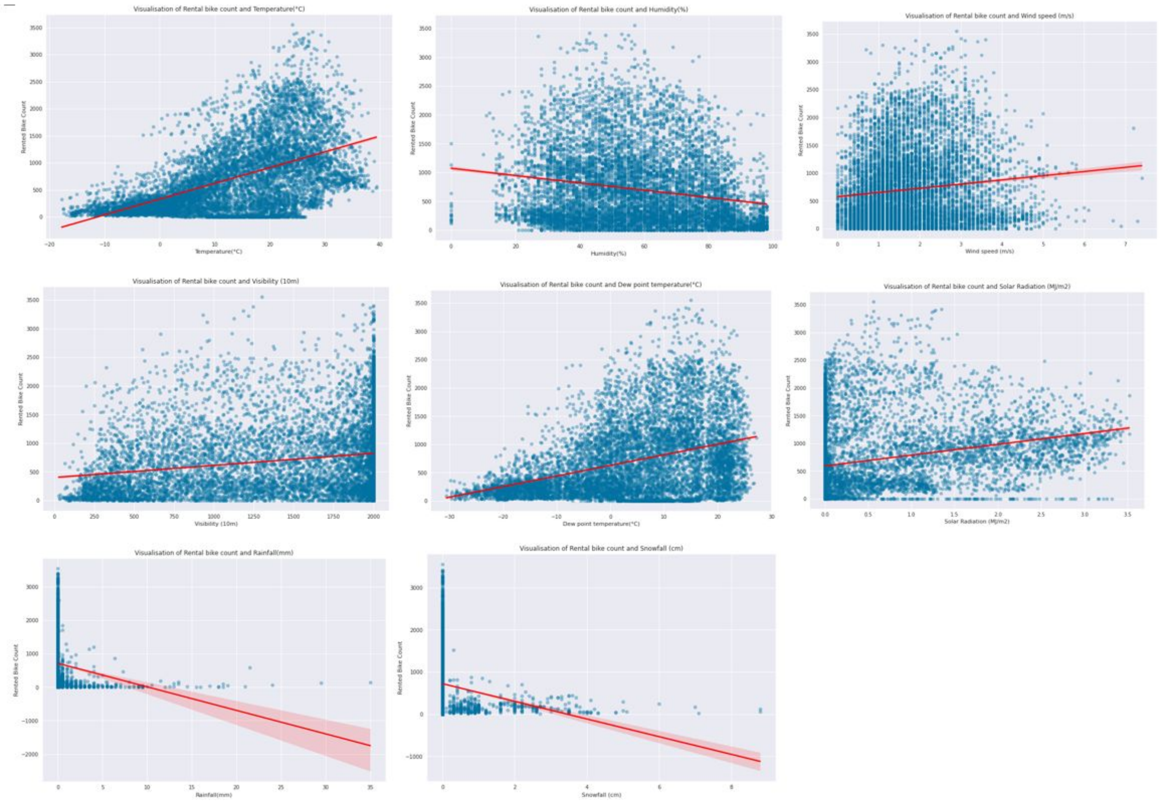


Data Analysis & Visualisation

- ❑ The scatterplot on the right shows the relationship between dependent variable and continuous variables.

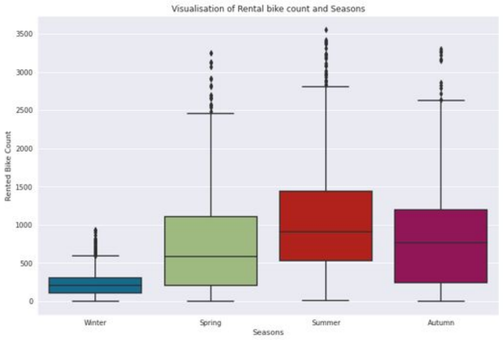
- ❑ It appears most of the continuous features have a positive linear relationship with the dependent variable.

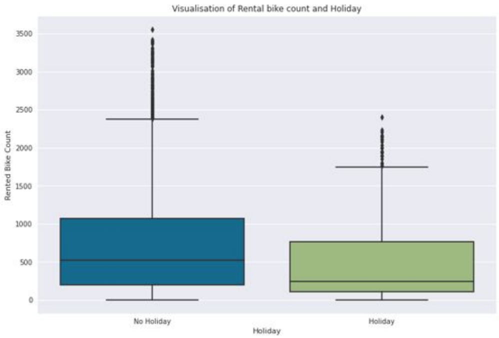
- ❑ Except for humidity, rainfall & snowfall, these have a negative linear relationship.

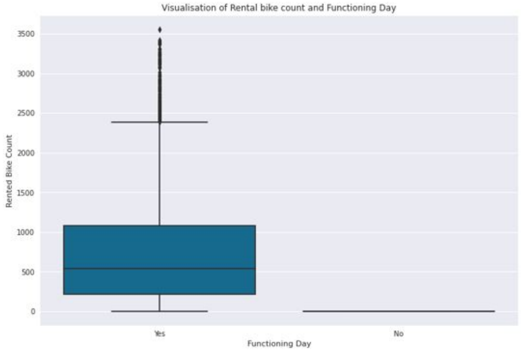


Data Analysis & Visualisation

- The relationship between the dependent variable and categorical variables is depicted via boxplots.
- We can deduce from the boxplot that there is a high demand for bicycles in the summer and autumn, but a low demand in the winter.
- There is very little demand for bicycles during the holidays.
- On non-functional days, bike demand is non-existent.


Visualisation of Rental bike count and Seasons


Visualisation of Rental bike count and Holiday


Visualisation of Rental bike count and Functioning Day

Data Analysis & Visualisation

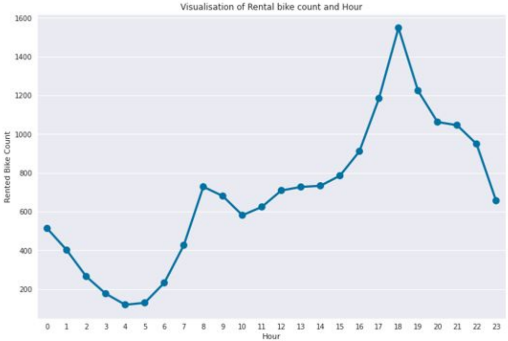
❑ The visualisations on the right depict the relation between dependent variable and cyclical features (time related).

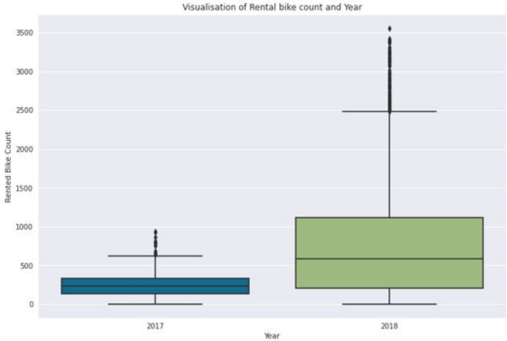
❑ It seems that there is a noticeable increase in the number of rented bikes between the 8th and 18th hour of the day.

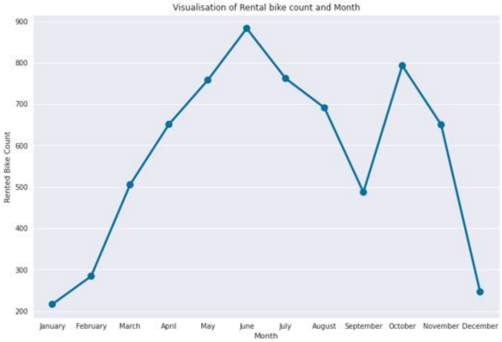
❑ As compared to 2017, the number of shared bikes on road climbed dramatically in 2018.

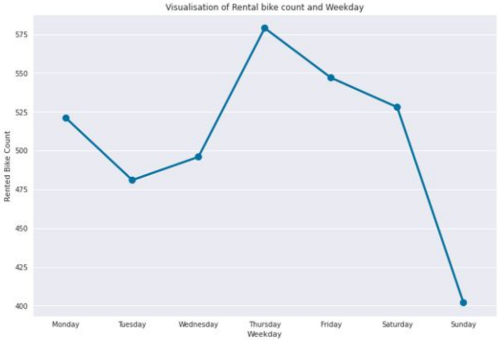
❑ In the months of May, June and July there is large demand for bikes.

❑ The demand for bicycles is highest on Thursdays and Fridays and lowest on Sundays.


Visualisation of Rental bike count and Hour


Visualisation of Rental bike count and Year


Visualisation of Rental bike count and Month

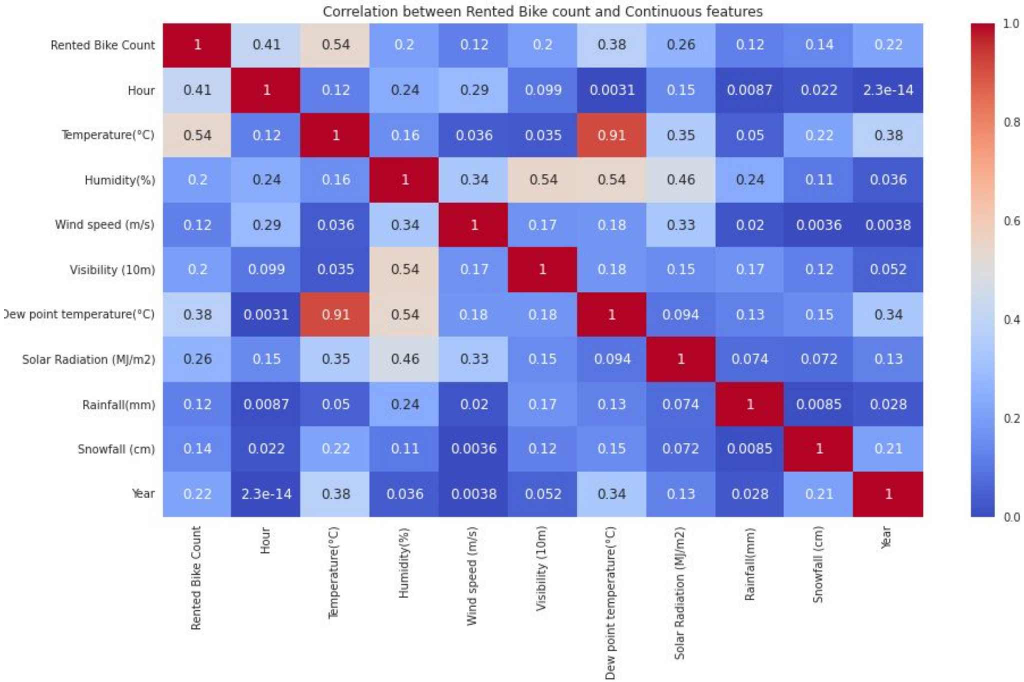

Visualisation of Rental bike count and Weekday

Data Analysis & Visualisation

- A heatmap on the right depicts the correlation between the continuous features and rented bike count.

- It shows that Hour and Temperature are highly correlated with Rented bike count.

- It also tells us multicollinearity is present in the data.



Correlation between Rented Bike count and Continuous features

	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Year
Rented Bike Count	1	0.41	0.54	0.2	0.12	0.2	0.38	0.26	0.12	0.14	0.22
Hour	0.41	1	0.12	0.24	0.29	0.099	0.0031	0.15	0.0087	0.022	2.3e-14
Temperature(°C)	0.54	0.12	1	0.16	0.036	0.035	0.91	0.35	0.05	0.22	0.38
Humidity(%)	0.2	0.24	0.16	1	0.34	0.54	0.54	0.46	0.24	0.11	0.036
Wind speed (m/s)	0.12	0.29	0.036	0.34	1	0.17	0.18	0.33	0.02	0.0036	0.0038
Visibility (10m)	0.2	0.099	0.035	0.54	0.17	1	0.18	0.15	0.17	0.12	0.052
Dew point temperature(°C)	0.38	0.0031	0.91	0.54	0.18	0.18	1	0.094	0.13	0.15	0.34
Solar Radiation (MJ/m2)	0.26	0.15	0.35	0.46	0.33	0.15	0.094	1	0.074	0.072	0.13
Rainfall(mm)	0.12	0.0087	0.05	0.24	0.02	0.17	0.13	0.074	1	0.0085	0.028
Snowfall (cm)	0.14	0.022	0.22	0.11	0.0036	0.12	0.15	0.072	0.0085	1	0.21
Year	0.22	2.3e-14	0.38	0.036	0.0038	0.052	0.34	0.13	0.028	0.21	1

Evaluation Metric

❑ The Root Mean Squared Error is the evaluation metric used in the project. It is the most widely used metric for assessing the accuracy of forecasts. It uses Euclidean distance to demonstrate how far predictions differ from true measured values.

❑ The root mean squared error is calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}},$$

where N is the number of data points, y(i) is the ith measurement, and ŷ(i) is its corresponding prediction.

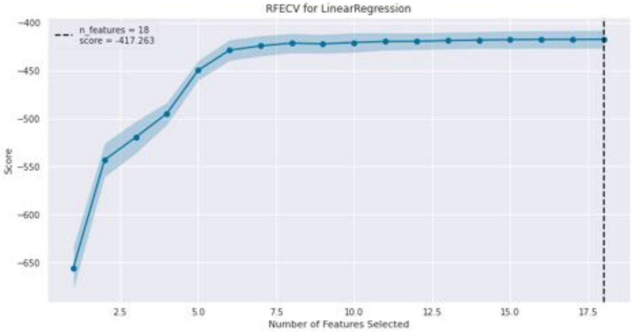
❑ In RMSE as the errors are squared before they are averaged, it gives a relatively high weight to large errors which makes RMSE most useful when large errors are particularly undesirable.

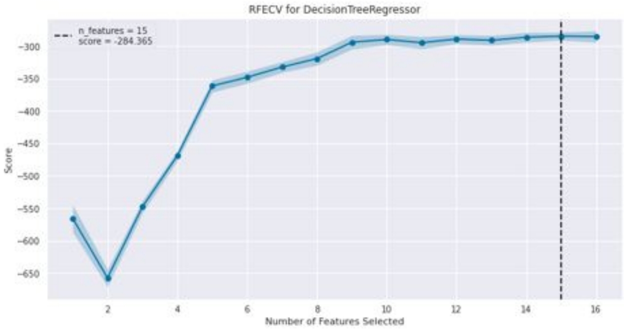
Feature Engineering

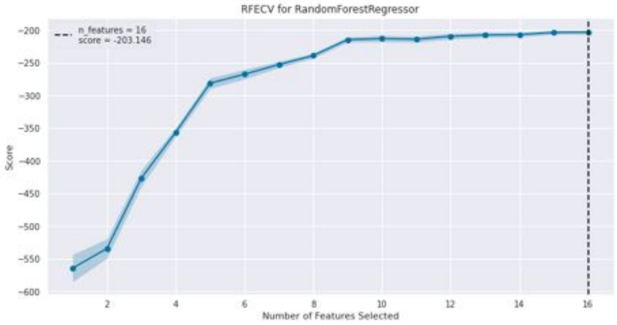
- Extracted new cyclical features from the date column and performed mathematical transformations.
- Detected multicollinearity using VIF, and dropped irrelevant features.
- MinMax scaling is used to transform the numerical variables.
- Applied square root transformation on dependent variable.
- For linear and distance-based models, one-hot encoding is used to transform the categorical variables into numerical representation, while for tree-based models, ordinal encoding is used for the transformation categorical variables.

Feature Selection

- ❑ Recursive Feature Elimination technique is used to select the best subset of features.
- ❑ Except for KNeighbors Regressor, all estimators with their default parameters selected their best features by wrapping themselves in RFE.



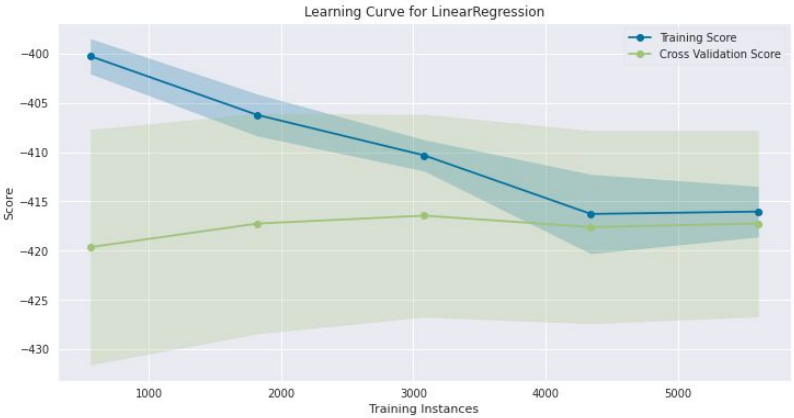




Model Implementation

Linear Regression:

- ❑ **Performance:**
  - ❑ RMSE on training set : 416.19
  - ❑ RMSE using cross-validation set: 417.26



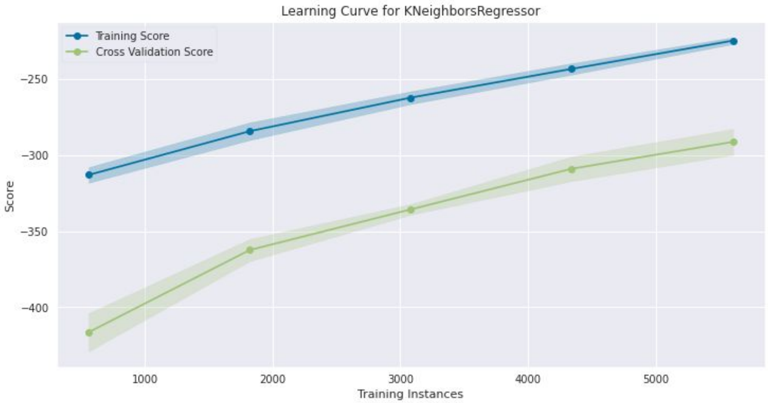
- ❑ **Remarks:**
  - ❑ Model has underfitted the data.
  - ❑ As I've already built new features and am using an unregularized model, the only way to improve performance would be to try more advanced models.

Model Implementation

KNeighbors Regressor:

- ❑ **Performance:**
  - ❑ **RMSE on training set : 207.02**
  - ❑ **RMSE using cross-validation set: 291.40**

- ❑ **Remarks:**
  - ❑ **Model underfitted the training data.**
  - ❑ **To reduce underfitting, the parameters must be hypertuned.**

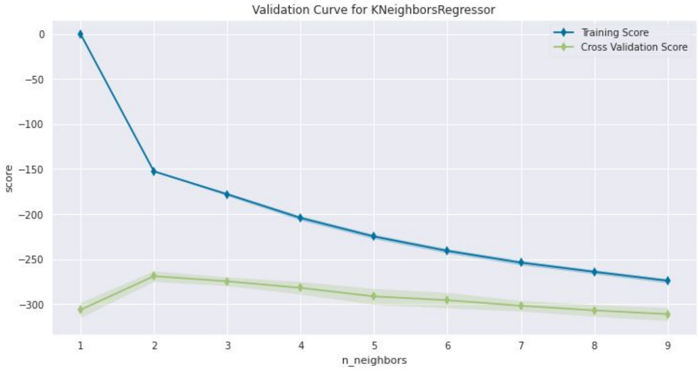


Learning Curve for KNeighborsRegressor

Model Implementation

KNeighbors Regressor:

- ❑ Hyperparameter tuning:
  - ❑ Best parameters:
    - ❑ n_neighbors : 2
  - ❑ Performance:
    - ❑ RMSE on training set : 143.18
    - ❑ RMSE using cross-validation set: 268.98

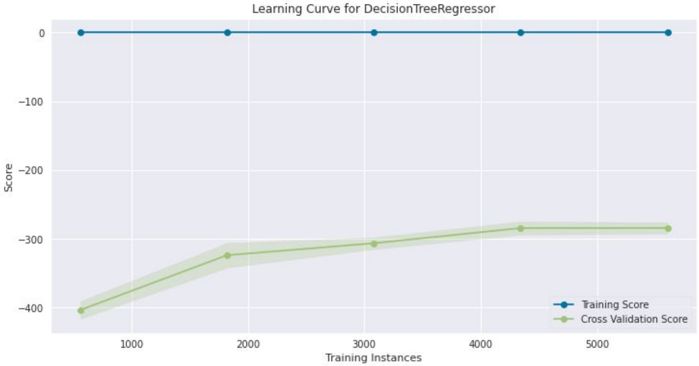


- ❑ Remarks after tuning:
  - ❑ Although hyperparameter tuning has improved the model's performance and performed better than Linear Regression, there is still room for improvement because the model is still underfitted.

Model Implementation

Decision Tree Regressor:

- ❑ **Performance:**
  - ❑ RMSE on training set : 0.0
  - ❑ RMSE using cross-validation set: 283.17



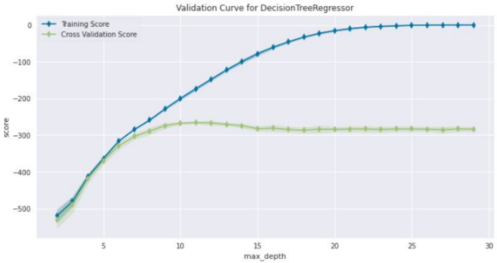
- ❑ **Remark:**
  - ❑ Model has overfitted the training data.
  - ❑ To regularise the model, the parameters must be hypertuned.

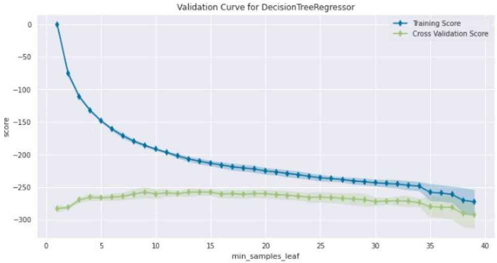
Model Implementation

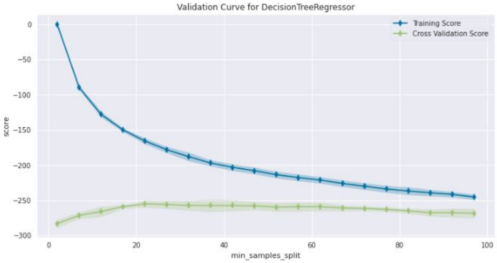
Decision Tree Regressor:

- ❑ Hyperparameter tuning:
  - ❑ Best parameters:
    - ❑ max_depth : 14
    - ❑ min_samples_leaf: 13
    - ❑ min_samples_split: 46
  - ❑ Performance:
    - ❑ RMSE on training set : 220.07
    - ❑ RMSE using cross-validation set: 256.77

- ❑ Remarks after tuning:
  - ❑ Although hyperparameter tuning improved the model's performance on the validation set, it has the opposite effect on the training set. As a result, when compared to KNeighbors Regressor, the model's overall performance falls behind.



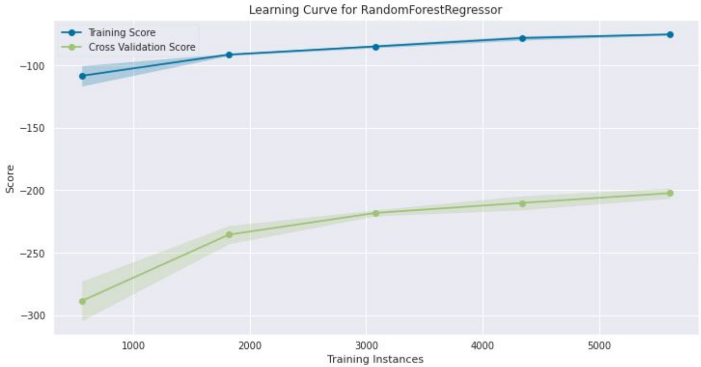




Model Implementation

Random Forest Regressor:

- ❑ Performance:
  - ❑ RMSE on training set : 73.69
  - ❑ RMSE using cross-validation set: 202.30

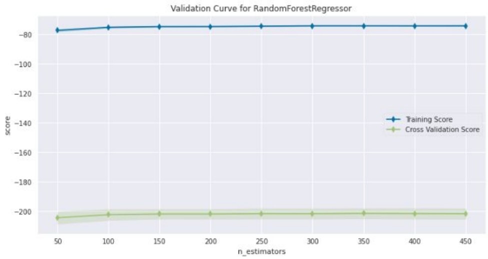


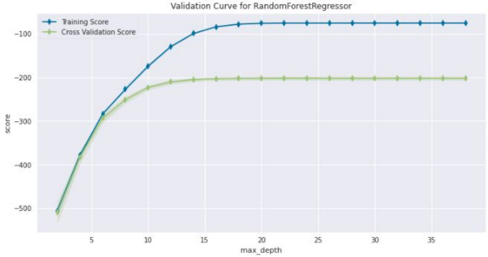
- ❑ Remark:
  - ❑ Model overfitted the training data.
  - ❑ To regularise the model, the parameters must be hypertuned.

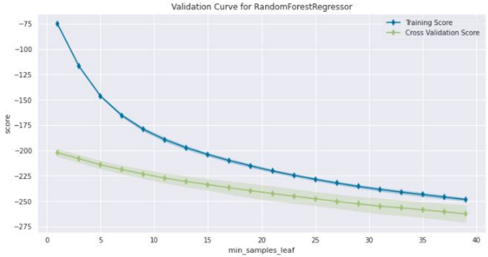
Model Implementation

Random Forest Regressor:

- Hyperparameter tuning:
  - Best parameters:
    - n_estimators : 350
    - max_depth : 14
    - min_samples_leaf: 13
  - Performance:
    - RMSE on training set : 72.24
    - RMSE using cross-validation set: 202.28
- Remarks after tuning:
  - There has been a minor improvement in the model. Despite the fact that the model is still overfitted, it outperforms all preceding models under current computational resources. As a result Random Forest is the best model.

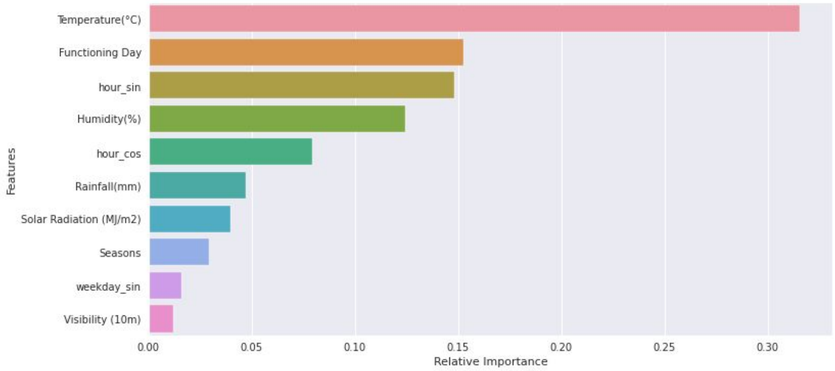

Validation Curve for RandomForestRegressor


Validation Curve for RandomForestRegressor


Validation Curve for RandomForestRegressor

Feature Importance & Final Model evaluation on test set

❑ The visualisation depicts the most important features on which bike count depends.



❑ The hypertuned Random Forest regressor has RMSE of 205.68 on the test set.

Challenges

❑ Feature Engineering of cyclical features.

❑ Selecting the range of hyperparameter values using validation curve graphs.

❑ Reducing the overfitting of the models.

❑ Tuning the models under current computational resources.

Conclusion

- ❑ As the temperature rises, the number of bikes increases, which could be useful for establishing new rental stations in hotter areas as well as maintaining bike supply in hotter areas.
- ❑ Companies renting bikes should maintain more supply on Functional days and can have maintenance of the bikes on non-functional days to decrease the down time in business.
- ❑ There is a high demand of bicycle during during 8th and 18th hour of the day, so there should be sufficient bikes to increase the business and the price of the bike could also be high during these hours.
- ❑ As bike demand rises in the summer and autumn seasons, prices should fall in the winter to increase use of bikes in winters.

Thank You