Bike Rental Analysis for BT Seoul Bike Hire

Presented by Yessy Rayner on 6 September 2022

Problem Definition

Since COVID-19 pandemic, bike rental businesses in Seoul are booming.

This is an unexpected result for BT Seoul Bike Hire, so they would like some analysis on:

- **How many extra staff members** require during busy period. It is estimated by the management that 1 staff member required to service every 200 bikes/customers
- **Quiet time of the day**, so they can service their bikes with less disruption. They are currently close 1 day per month to service their bikes
- BT Seoul Bike Hire also would like to **expand their business to other major metropolitan cities** such as Busan and Incheon using the same business model and staffing method.

Dataset Description

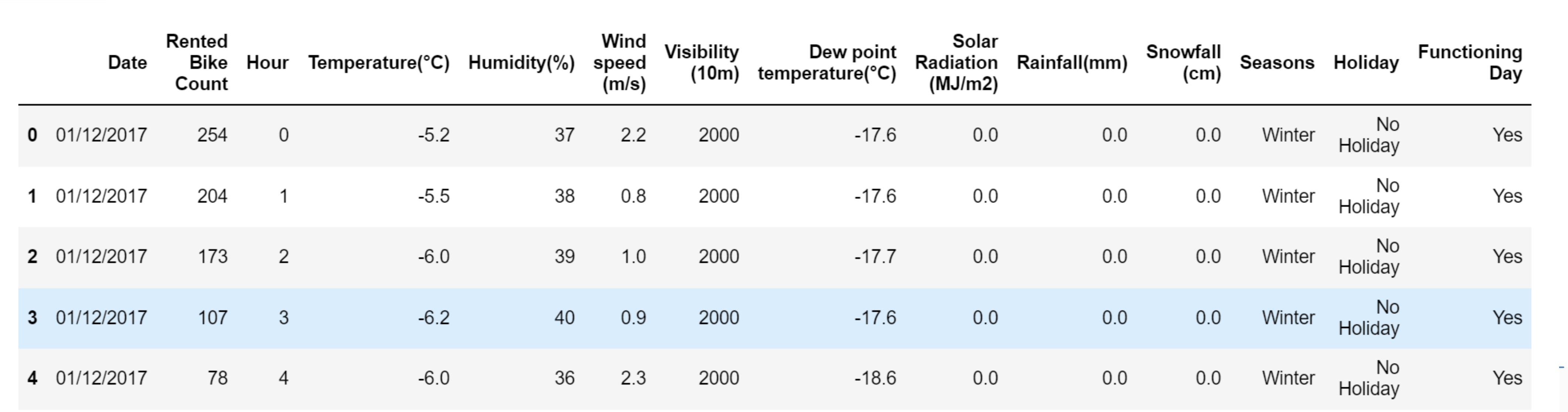
- The dataset contains rented bike count at each hour with the corresponding weather data and date information for the six months period (From March to October 2018).
- **Over 7890 rented bike data with 10 features** – which will be group as follow:

  Base variables – Temperature (C), Dewpoint (C), Solar radiation (MJ/m2) due to apparent linear relationship

  Continuous variables – Humidity (%), Windspeed (m/s), Visibility (10m), Snowfall (cm) and Rainfall (mm)

  Discrete variables – Date and Hour

  Categorical variables – Seasons and Holiday



	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

Preprocessing / Data Cleaning

Here are some of the data preprocessing and cleaning methods being performed:

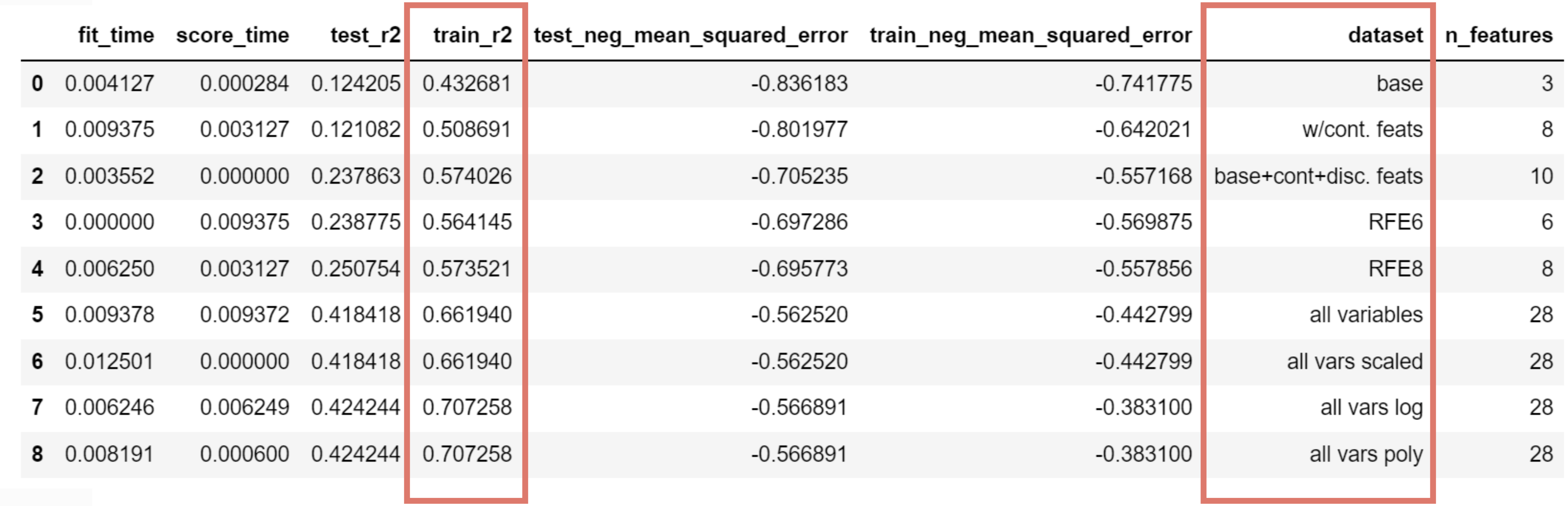
- Checked if all data are in **correct data types**
- Checked any **missing data**
- Added 2 additional categorial variables based on:
  Date → **Day_of_week:** Monday to Sunday
  Hour → **Shift:** 4x Shift based on 6 hour per shift (Early morning, Morning peak, Mid day, Evening peak) – This method will help with predicting the staffing requirements
- Drop 'Functioning Day = No' (Close down due to bike servicing day)
  Removed the 'No' value as it skews the overall data.

Modelling

- Using **linear regression model** to predict bike rental per hour or shift based on numerous features/variables
- **Cross validation -** where data being split into train and test data (75/25 split)
- **Testing on different variables/models based on:**
  Model 1: Base variables
  Model 2: Base + continuous variables
  Model 3: Base + continuous + discrete variables
  Model 4: Recursive Features Elimination
  Model 5: ALL variables including categorial variables after being transformed using OneHotEncoder
- **Popular Transformation methods** performed in order to get the best fit → Logarithm, Min-Max Scaling, Standard Scaling, Polynomial-Interaction.

Result 1

- **R-squared score at 0.707 (Moderate Linear Relationship)**
  R-squared shows how well the data fit the regression model (the goodness of fit).
  1 indicated the perfect fit/perfect score.

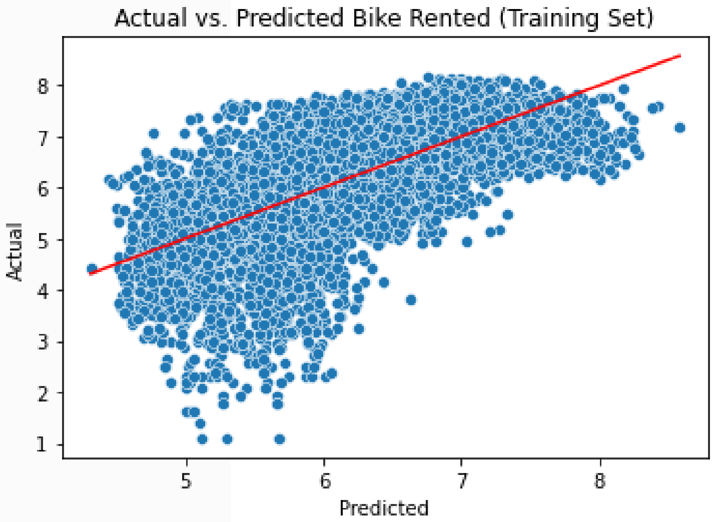


	fit_time	score_time	test_r2	train_r2	test_neg_mean_squared_error	train_neg_mean_squared_error	dataset	n_features
0	0.004127	0.000284	0.124205	0.432681	-0.836183	-0.741775	base	3
1	0.009375	0.003127	0.121082	0.508691	-0.801977	-0.642021	w/cont. feats	8
2	0.003552	0.000000	0.237863	0.574026	-0.705235	-0.557168	base+cont+disc. feats	10
3	0.000000	0.009375	0.238775	0.564145	-0.697286	-0.569875	RFE6	6
4	0.006250	0.003127	0.250754	0.573521	-0.695773	-0.557856	RFE8	8
5	0.009378	0.009372	0.418418	0.661940	-0.562520	-0.442799	all variables	28
6	0.012501	0.000000	0.418418	0.661940	-0.562520	-0.442799	all vars scaled	28
7	0.006246	0.006249	0.424244	0.707258	-0.566891	-0.383100	all vars log	28
8	0.008191	0.000600	0.424244	0.707258	-0.566891	-0.383100	all vars poly	28

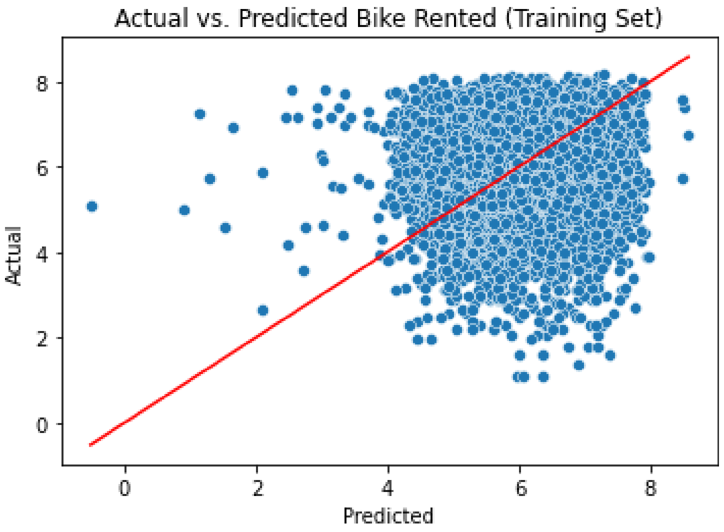
Result 2

Linear regression fit models improvement

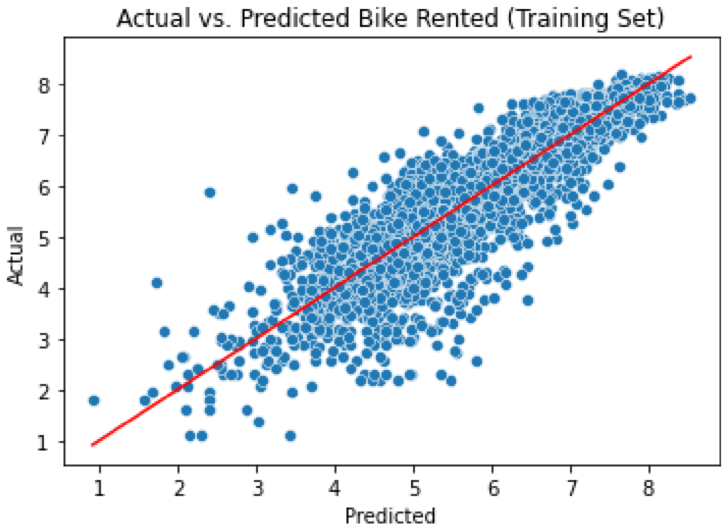
Base variables



ALL variables

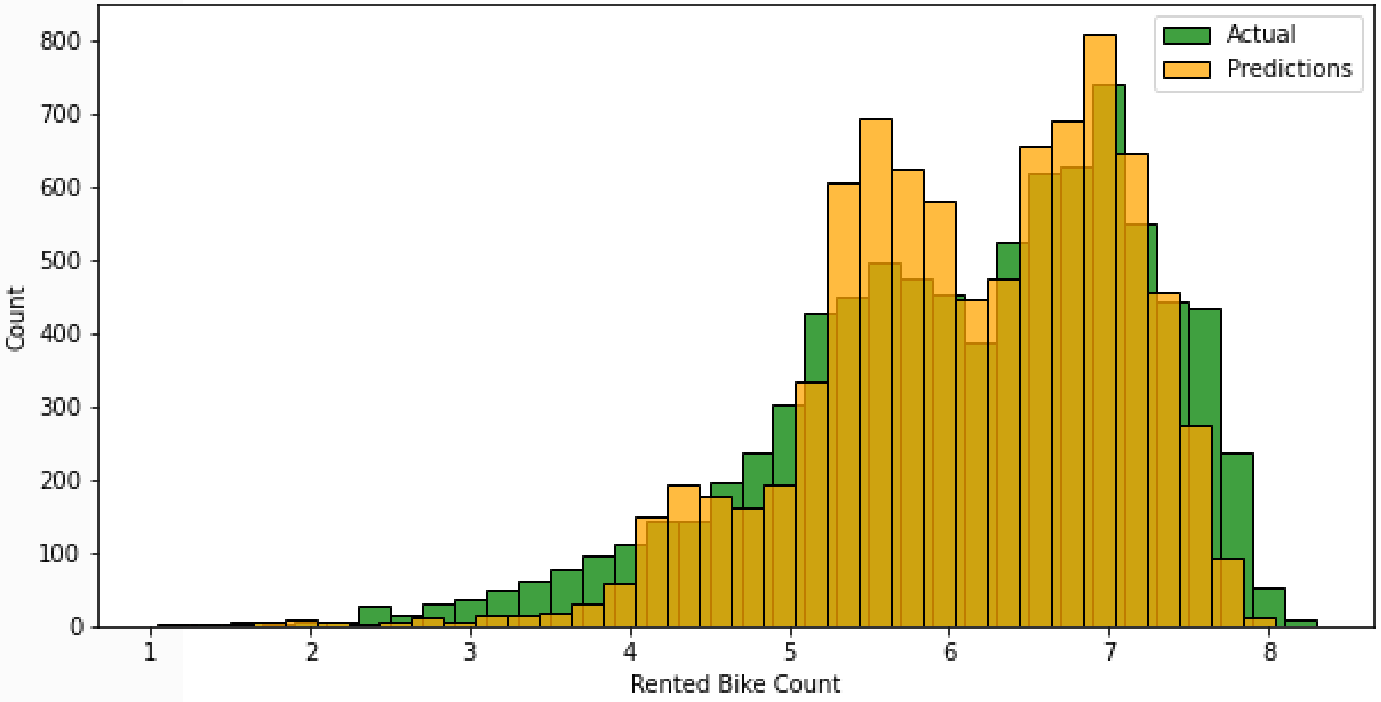


Final model



Result 3

Compare ranges of prediction to actual values

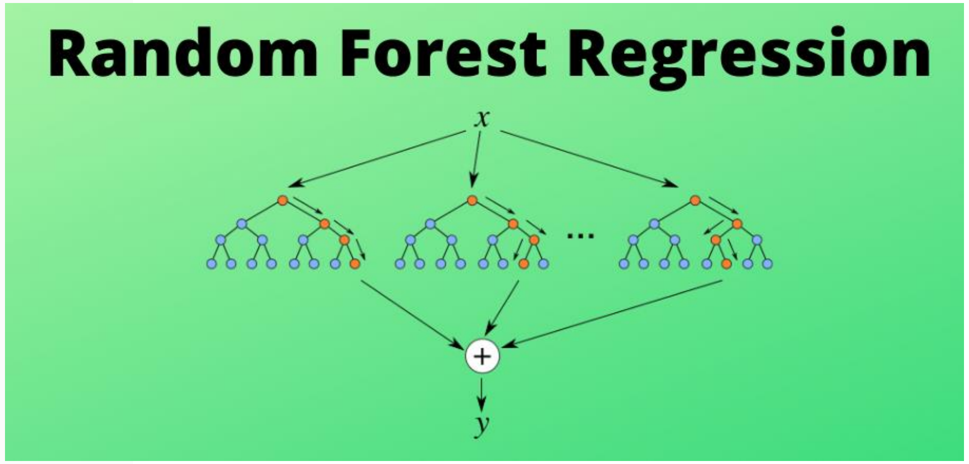


Most important features:

- Temperature
- Humidity
- Visibility
- ALL of the categorical variables e.g. Day of Week, Seasons, Shift

Next Step

- Gather more bike rental data especially post COVID-19 period, ideally throughout 2022 to allow for more accurate prediction
- Try **Random Forest Regression** model might work well with the Bike Rental data



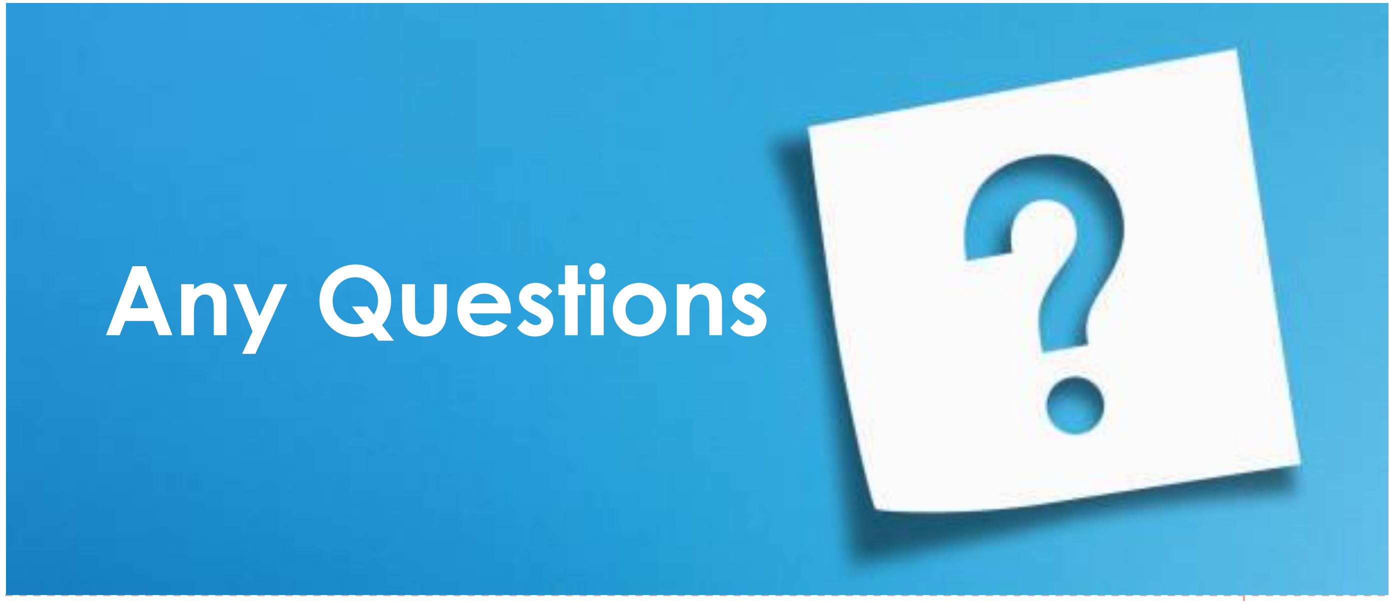
Limitation

- R-squared score of 0.707 is not perfect despite numerous models and transformation being tested
- However, 0.707 is pretty good score especially in predicting human behaviour
- In business/staffing modelling, 0.707 is adequate as it provides moderate linear relationship

Random Forest Regression is a prediction model based on the trees structure and it takes into account of many predictions.
This is because of the average value used. These algorithms are more stable because any changes in dataset can impact one tree but not the forest of trees.

Thank you!

Presenter:	Yessy Rayner
Email:	Yessy.Rayner@gmail.com
LinkedIn:	https://www.linkedin.com/in/yessy-rayner
GitHub:	https://github.com/YessyLee



Any Questions