DATA 245: MACHINE LEARNING TECHNOLOGIES

Bike Sharing Demand Prediction

Image source: tenor.com

LOHITHA VANTERU, MAHE JABEEN ABDUL,
PRANAVI SANDRUGU, DIVYA NALAM

TODAY'S HIGHLIGHTS

- What is a bike-sharing system? What is the issue?
- Introducing Bike Sharing Rebalancing
- Observing Patterns
- Data
- Our approach
- When are the most rides happening?
- What is the distribution of rides on different kinds of days w.r.t .Hours?
- What is the distribution and correlation among variables ?
- Data Transformation
- Model Development
- Solving the issue with Reinforcement Learning.
- Avoiding overfitting
- Evaluation
- Conclusion and Future Works

Discussion Outline

what is a bike-sharing system?

- ONE OF THE POPULAR MICROMOBILITY SERVICES
- IT'S FASTER THAN A CAB AND LESS CROWED THAN THE SUBWAYS.
- IT'S GREEN!

But, the problem is: Getting these bikes at popular locations is highly impossible.

Image Source: : https://www.marketplace.org/2021/10/06/bike-share-programs-arent-profitable-but-chip-away-at-emissions/

Image Source: https://brooklyneagle.com/articles/2017/12/18/nyc-success-of-city-wide-bike-share-program-not-wide-enough/

Image Source: https://www.itdp.org/2018/01/25/regulating-dockless-bikeshare/

It's either no bikes at all or an absolute clutter.

But, we have a solution.

Introducing Bike Sharing Rebalancing

How?

By predicting the bike sharing demand

Using?

Linear regression models like Lasso, Ridge, Polynomial, and Elastic regressions, and tree-based models like Decision Tree, Random Forest, and Gradient Boosting.

Observing Patterns

WHO EVEN RETURNS A BIKE TO A STATION ON A HILLTOP WHEN IT'S HOT?

The above example is one such pattern.

Image Source: https://drawception.com/game/tPPzDm86wa/biking-up-a-hill/

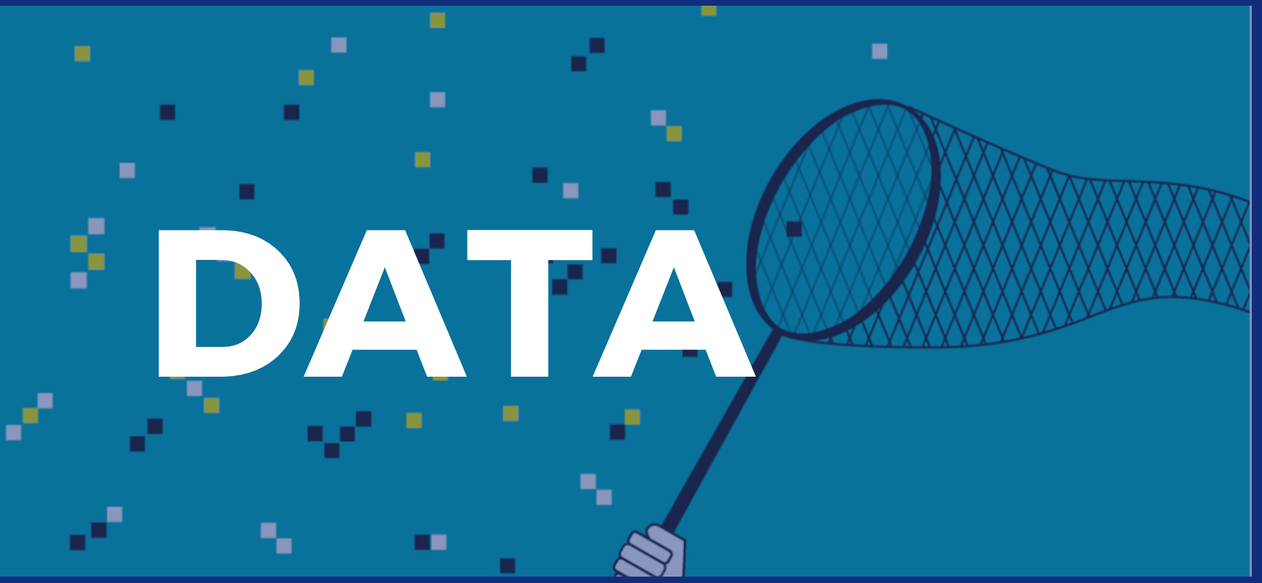


Image Source: https://hbr.org/2017/06/does-your-company-know-what-to-do-with-all-its-data

But, we need data to observe patterns.

We used data from the Machine learning repository which has details about hourly bike rental counts during the period of 2017 and 2018 along with the weather information during that time.

Attribute Information:

- Date : year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of he day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10m
- Dew point temperature - Celsius
- Solar radiation - MJ/m2
- Rainfall - mm
- Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

Attribute Information

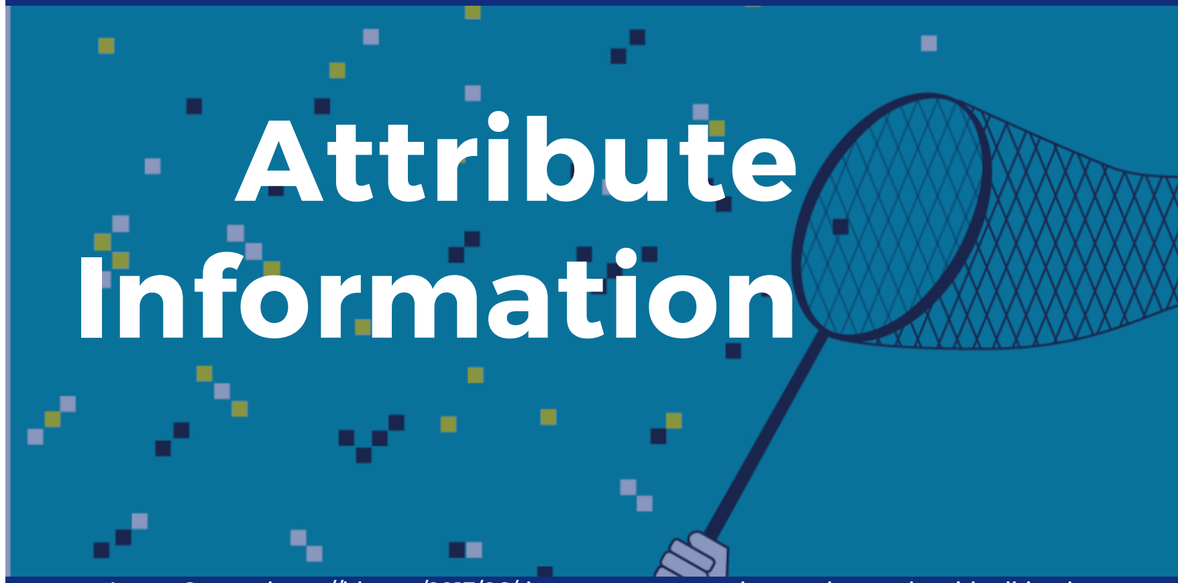
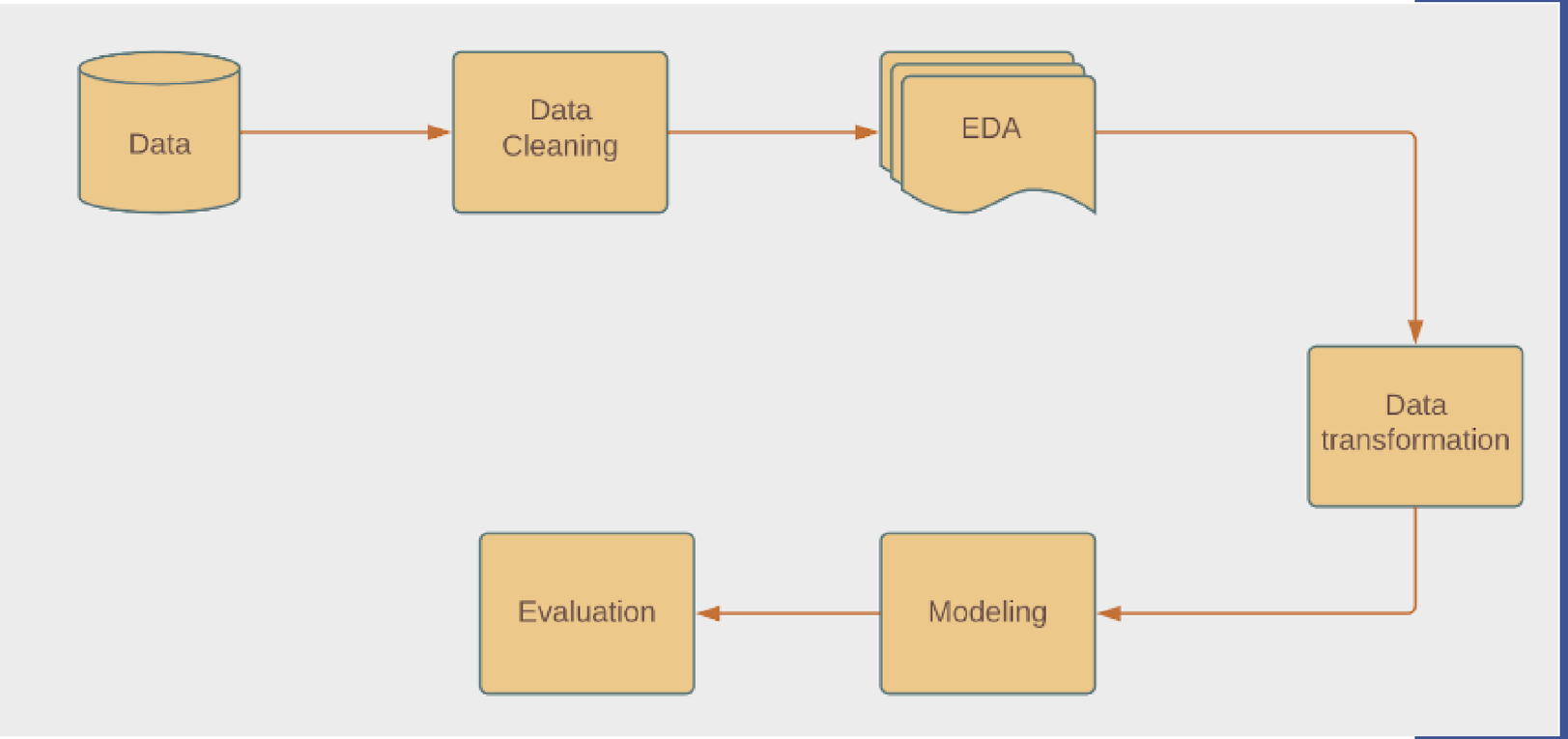


Image Source: https://hbr.org/2017/06/does-your-company-know-what-to-do-with-all-its-data

Our Approach



WHEN ARE THE MOST RIDES HAPPENING?

We performed Exploratory Data Analysis to understand when the demand is high.

Imager Source: https://www.questionpro.com/blog/exploratory-data-analysis/

Observations:

- **Time of day**: It is evident that the peaks are observed in the mornings at 8 AM & 9 AM and evenings at 4 PM, 5 PM & 6 PM implicating that the majority of the trips are taken before and after the usual office hours. Moreover, the trips taken in between the usual office hours i.e. 10 AM — 3 PM are constant. Additionally, the trips have declined substantially after 6 PM.
- **Day of week:** As the majority of the trips were taken before and after the usual office hours, it is no surprise that the majority of the trips have taken place on weekdays (Mon-Fri) as compared to weekends (Sat-Sun)
- **Season:** We observe that summer has the highest rentals followed by spring and then fall, which gives us reason to believe that bike riders prefer warm to pleasant climates than colder climates.

Image Source:https://www.freepik.com/vectors/boy-thinking

WHAT IS THE DISTRIBUTION OF RIDES ON DIFFERENT KINDS OF DAYS W.R.T. HOURS?

Observations :

- We can observe that the pattern of weekdays and weekends as well as Holidays is different, in the weekend and on holiday the demand becomes high in the afternoon. While the demand for office timings is high during weekdays and no holidays.
- For Month, we can clearly observe that the demand is low in December, January & February, as it is cold in these months and we have already seen w.r.t. season that demand is less in winters.
- Comparatively the demand was higher in 2018 than in 2017, which can support the fact that bike share demand is increasing over the years.

Image Source:https://www.freepik.com/vectors/boy-thinking

What is the distribution and correlation among variables ?

Observations :

- From the distplots, we could see that some of the variables are either right or left skewed and for those variables, even the mean and median are skewed as seen in the histograms.
- Regression plots show that a few variables are positively correlated and few are negatively correlated with our target variable.
- From the Heat map, we could see that there is multicollinearity between a few variables like Temperature and Dew point temperature.
- From the pie plots, we could see that the data is uniformly distributed across all seasons, months and hours. Hence, there is no data imbalance.

Image Source:https://www.freepik.com/vectors/boy-thinking

DATA TRANSFORMATION

- To apply linear regression models, we checked the multicollinearity across all independent variables and dropped the variables like Dew Point Temperature which has a variable inflation factor greater than five.

- We have used label encoding to transform the categorical variables to get dummy variables. We observed that the distribution of the target variable is skewed and hence applied logarithmic and square root transformation on it and after checking the skewness factor, we decided to continue with square root transformation on the target variable for linear regression modeling.

- Furthermore, based on our exploratory data analysis, we have dropped the features which are not important and proceeded with the features which have more correlation with the dependent variable.

- As Decision trees make no assumptions about relationships between features, we used all the independent features because multicollinearity won't affect the model.

- Also, we didn’t transform the target variable here, as its distribution won't have an impact on model accuracy. We have scaled the features to be normalized and have a single unit variance.

- Our Project is to predict hourly bike share demand at a particular location on a particular day. This is considered as a regression problem as the bike rental count is a continuous value.

- The association between input features and the continuous output variable is determined using regression techniques, a variety of machine learning algorithms, like Linear regression models including Lasso, Ridge, Polynomial and Elastic regressions, and tree-based models like Decision Tree, Random Forest, and Gradient Boosting.

- Using the train test split function, we divided the dataset as 75% for training and 25% for testing and got the values for the train and test sets. We implemented the models using the X train, y train, X test, and y test variables returned from the train test split function.

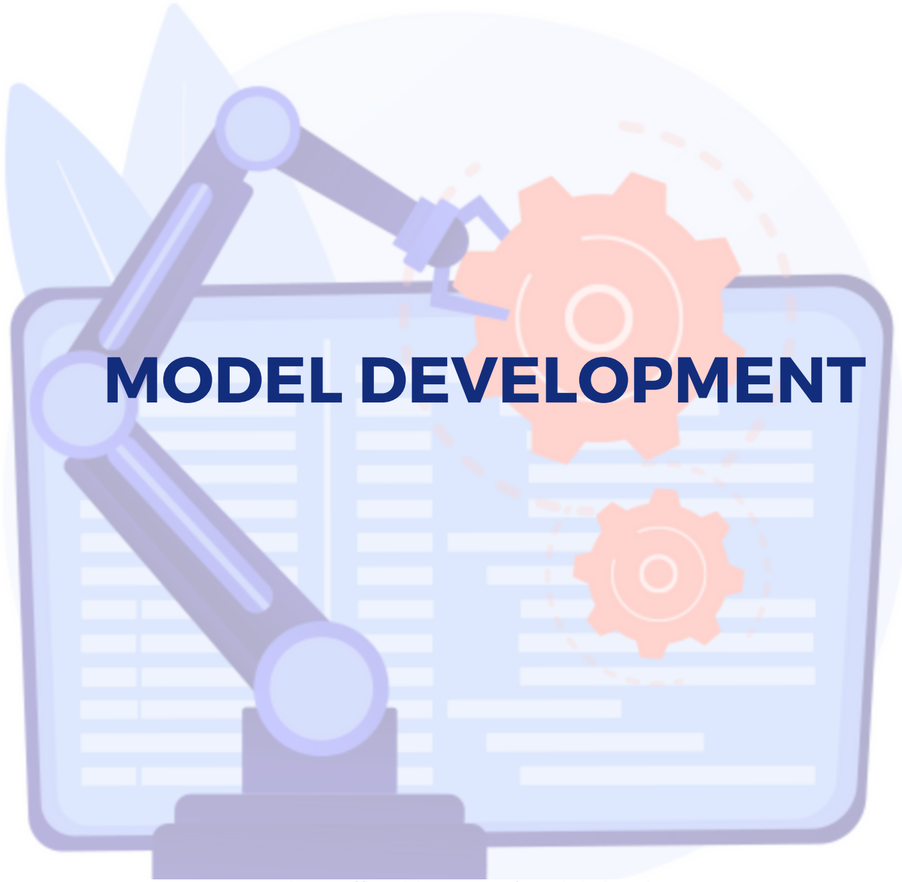


Image Source: https://hackient.com/model-development

Solving the issue with Reinforcement Learning

The reward and punishment structure are as follows:

- -30 if the hourly bike stock falls outside the range [0, 100].
- +20 if bike stock is in the range [0, 100] at 23 hours; otherwise, -20
- -0.5 times the number of bikes eliminated every hour.

This rewards scheme incentivizes the agent to move as few bikes as possible while maintaining the bike supply within a reasonable range.

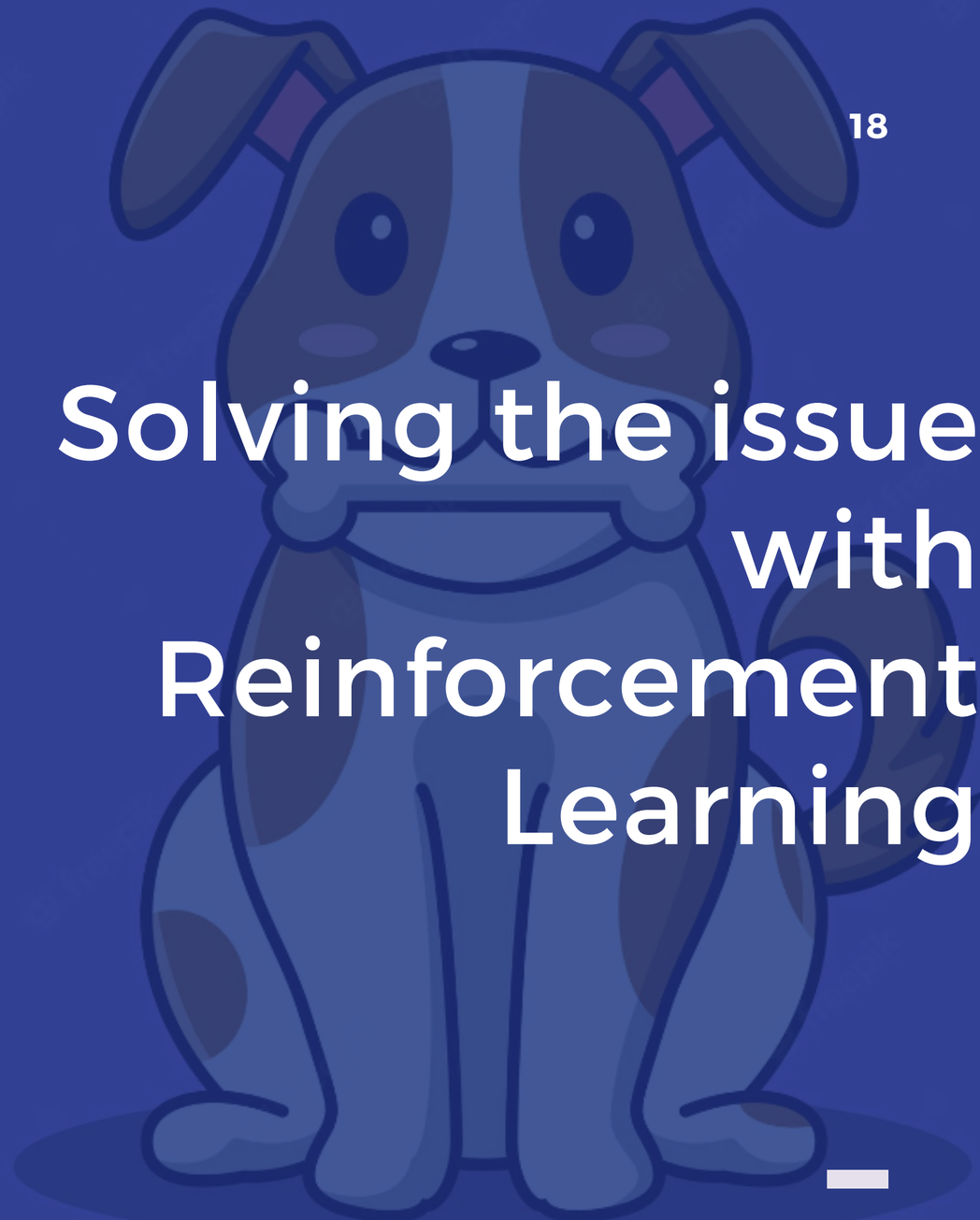


Image Resource: https://www.freepik.com/premium-vector/cute-dog-cartoon-eating-bone-vector-cartoon-illustration-cartoon-clipart_26014075.htm

Avoiding overfitting

To avoid overfitting, each model must be fine-tuned using its optimal hyperparameters. To identify the ideal hyperparameters, a grid search method with repeated cross-validation (CV) was employed.

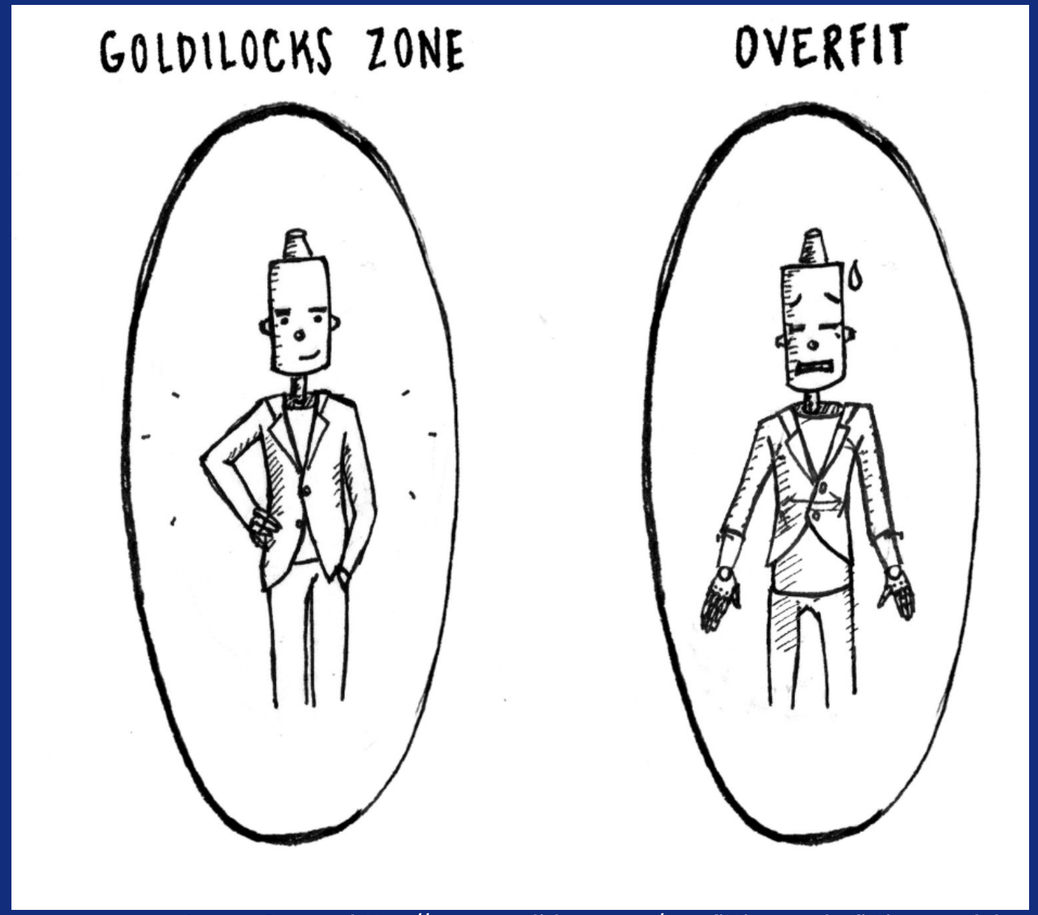


Image Source: https://www.euclidean.com/overfitting-underfitting-models

EVALUATION

We are using the following metrics :

- **Mean Squared Error(MSE)**
- **Root Mean Squared Error(RMSE)**
- **Mean Absolute Error(MAE)**
- **Accuracy**
- **R2 and adjusted R2**

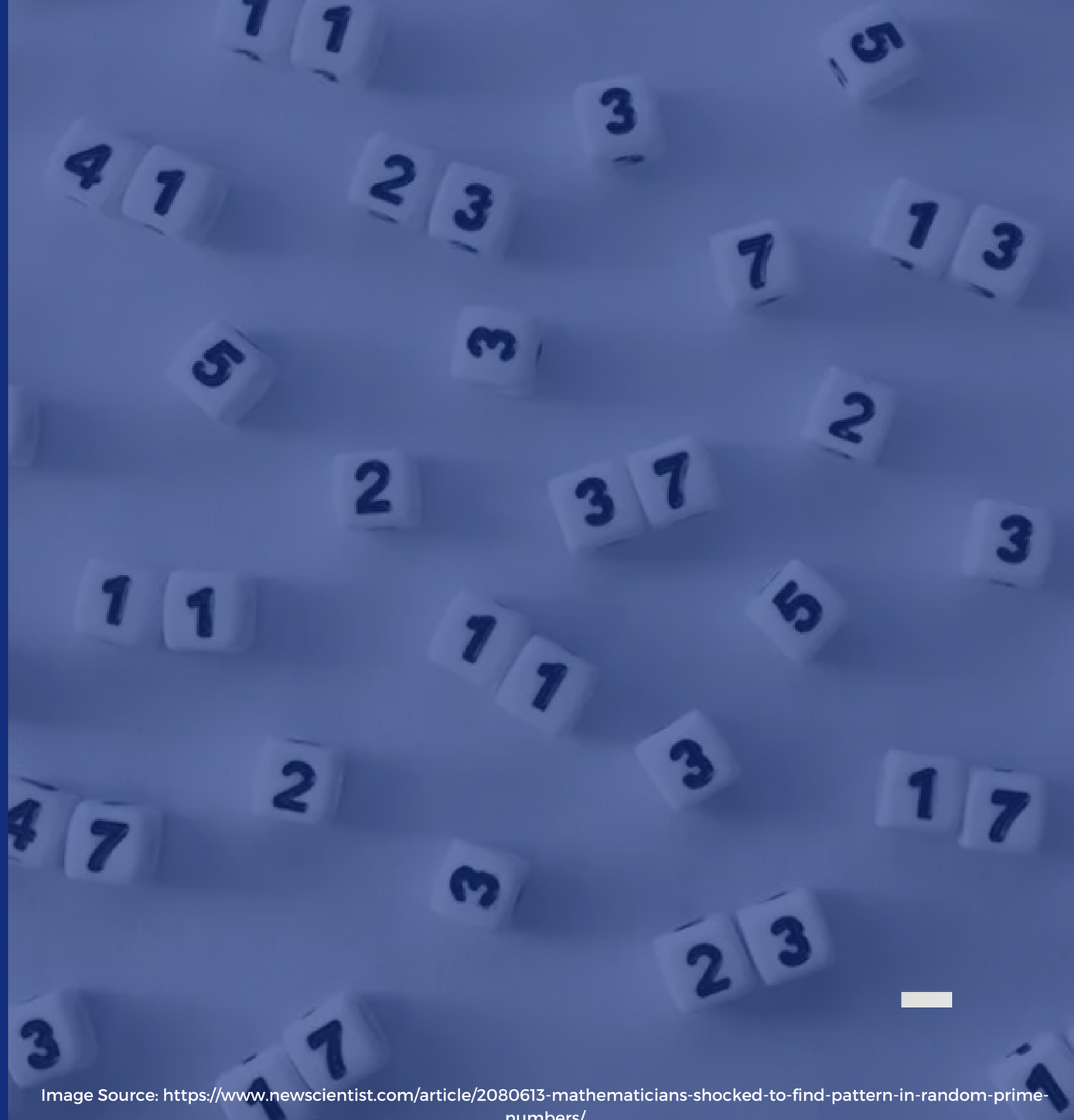


Image Source: https://www.newscientist.com/article/2080613-mathematicians-shocked-to-find-pattern-in-random-prime-numbers/

	model	Mean_Absolute_error	Mean_square_error	Root_Mean_square_error	Train_Accuracy_score	Test_Accuracy_score	R2	Adjusted_R2
)	Random_Forest	261780.893059	3.364247e+11	580021.293845	0.982798	0.882351	0.823853	0.819819
l	xg_boost	360262.242007	5.094854e+11	713782.441591	0.830808	0.820552	0.733240	0.727132
2	Linear Regression	225.127202	1.107577e+05	332.802775	0.759672	0.759672	0.728831	0.722752
3	Lasso	225.138710	1.107702e+05	332.821621	0.759672	0.759672	0.728801	0.722720
4	Decision_Tree	348790.397260	5.779244e+11	760213.372079	1.000000	0.785320	0.697407	0.690478
5	Ridge	301.417177	1.987573e+05	445.822076	0.627804	0.627804	0.513381	0.502471
6	Elastic	303.385845	1.988263e+05	445.899441	0.618005	0.618005	0.513212	0.502299

METRICS BEFORE HYPERPARAMETER TUNING

model	Mean_Absolute_error	Mean_square_error	Root_Mean_square_error	Train_Accuracy_score	Test_Accuracy_score	R2	Adjusted_R2
xg_boost	139.845106	45133.130162	212.445593	0.880039	0.849289	0.885042	0.882410
Random_Forest	186.118814	77467.485443	278.329814	0.808248	0.784661	0.802683	0.798165
Decision_Tree	174.241781	89442.409018	299.069238	0.789514	0.679515	0.772182	0.766966
Linear Regression	225.127202	110757.687257	332.802775	0.754873	0.737842	0.728831	0.722752
Polynomial	225.127202	110757.687257	332.802775	0.754873	0.737842	0.728831	0.722752
Ridge	225.171103	110812.502363	332.885119	0.754847	0.738251	0.728697	0.722615
Lasso	225.377829	111014.518473	333.188413	0.754909	0.742090	0.728202	0.722109
Elastic	225.756265	111412.146780	333.784581	0.754848	0.739109	0.727229	0.721114

METRICS POST HYPERPARAMETER TUNING

KEY LEARNINGS

- Comprehending the existing research on bike-sharing systems and coming up with an approach to work on our goals, and selecting the machine learning models to carry on our work is the best experience for everyone on our team.

- We were able to work together whenever necessary and shared our thoughts and responsibilities to proceed with our goals, and managed our team meetings wisely, which helped us advance our managerial abilities.

- A crucial step in modeling any machine learning system is choosing the features required for prediction. Whether using supervised or unsupervised techniques, picking the best machine learning methods for the data set is crucial.

- After collecting the data, we must clean the data before modeling; otherwise, modeling will fail and will be time-consuming work to clean and re-fit the model.

- We were able to execute EDA in a significant way which improved our understanding of the distribution of bike riders data and helped us perform different types of visualizations.
- We observed that tree-based models require less data preparation compared to linear models as it doesn't require normalized data or scaling of features.

- With the available evaluation metrics for regression, we selected the metrics that predominantly helped us learn and gain more insights. We were able to understand the applicability of the models based on the context by using several machine learning models.

- We understood the importance of tuning the model and a good choice of hyperparameters can really improve the algorithm performance.

- Model explainability is extremely useful to understand the models and debug them in order to make informed decisions on how to debug them.

CONCLUSION AND FUTURE WORKS

CONCLUSION

We observed how different parameters like weather, day of the week and others impact the output. The performance of the models can be impacted by hyperparameters, which are critical to the outcome of machine learning algorithms. To avoid overfitting, each model must be fine-tuned using its optimal hyperparameters.

FUTURE WORKS

We want to extend our scope of work to enhance our chances of publishing, with the additional features of prioritizing bike access to healthcare workers or emergencies, providing the option to reserve the bikes beforehand for their convenient time slots, and keeping the rebalancing cost as low as possible. As an extension to this work, we would like to use any cutting-edge deep learning technologies to predict the demand for bike sharing and to solve the rebalancing issues.

THANK YOU!

Any questions?