Exploring Asteroid Diameter : Predictive Modeling for Celestial Bodies

Aditya Shanmugham, Jonathan Paul Dande, Srindhi Sunkara

1. Summary

The project aims to leverage advanced AI algorithms to improve the accuracy of asteroid diameter predictions, thereby contributing to our understanding of asteroid dynamics and bolstering our capacity to mitigate potential threats to Earth. The impact of a large asteroid could lead to devastating environmental consequences like global climate change and tsunamis. Early detection of such threats allows for the possibility of deflection methods. Asteroid diameters are critical to understanding their dynamical and morphological evolution, potential as spacecraft targets, impact threat, and much more, yet most asteroid diameters are uncertain by >50 per cent because of the difficulties involved in calculating diameter due to their irregular shapes [2].

By delving into the factors influencing the accuracy of asteroid diameter predictions, the project endeavors to harness historical asteroid data from the Jet Propulsion Laboratory's (JPL) database. This dataset contains 26 columns and over 839k rows, encompasses a wealth of asteroid-related information, including orbital parameters, physical traits, and known diameter measurements.

Throughout the project, we performed a comprehensive analysis of the dataset to unravel the factors influencing asteroid diameter predictions. To achieve this, we did various preprocessing steps including handling missing values, label encoding, normalizing and removing outliers. We utilized correlation mapping and Pearson coefficient analysis to visualize and quantify the relationships between features and the target variable, aiding in the identification of highly correlated features essential for predicting asteroid diameter accurately. We found the top performing algorithms from a list of 25 ML algorithms using LazyPredict and then we fine-tuned the top performing algorithms in a comparison nature and understood the reasons for the results obtained.

This detailed analysis enabled us to discern patterns in asteroid characteristics and diameter measurements, providing valuable insights into asteroid dynamics and compositions. We found that certain types of models, like LightGBM, were better at making accurate predictions compared to others. LightGBM, in particular, stood out for its ability to make predictions that were very close to the actual asteroid diameters. This means that it could be a valuable tool for accurately estimating asteroid sizes in the future.

2. Methods

2.1 Programming Language: Python

2.2 Libraries Used: pandas, numpy, sklearn, seaborn, XGboost, matplotlib, scipy

2.3 Pre Processing:

1. Removed missing or null values in the dataset
2. Removed rows without information about "diameter."
3. Imputed information into missing data
4. Identified and removed any outliers
5. Converted categorical columns to numerical type

6. Identified less correlated features w/target variable and the applied PCA
7. Divided the dataset into training, validation, and testing sets for model evaluation

2.4 Exploratory Data Analysis

The target ('diameter') variable is missing 83% of the data (696k out of 839k). So, our first step is to remove all the rows where the target variable is missing. After this filtering, we lost about 83% of data which leaves us with approx 137k of rows.

The dataset has a total of 24 features out of which some of the features have more than 90% of missing data. So we filtered out such features if the percentage of missing data is more than 60%. Some of the columns which were affected : ***'G', 'extent', 'rot_per', 'GM', 'BV', 'UB', 'IR', 'spec_B', 'spec_T'.*** After this step, we are left with 15 features.

For the features with less than 1% of missing values, we imputed the missing values with the mean of the feature set. ***'Data_arc', 'H', 'albedo'*** were some of the features which required this imputation.

Once the missing values are handled, We cross referenced the features from the data with the data documentation found in NASA's official website. We looked for any mismatch of data types, invalid data, and potential outliers. We deemed the data points to be outliers if they fall outside the range of ***1.5 times of IQR from Q1 and Q3.*** This resulted in the below box plots for the data before cleaning (Figure 1) and after cleaning (Figure 2)

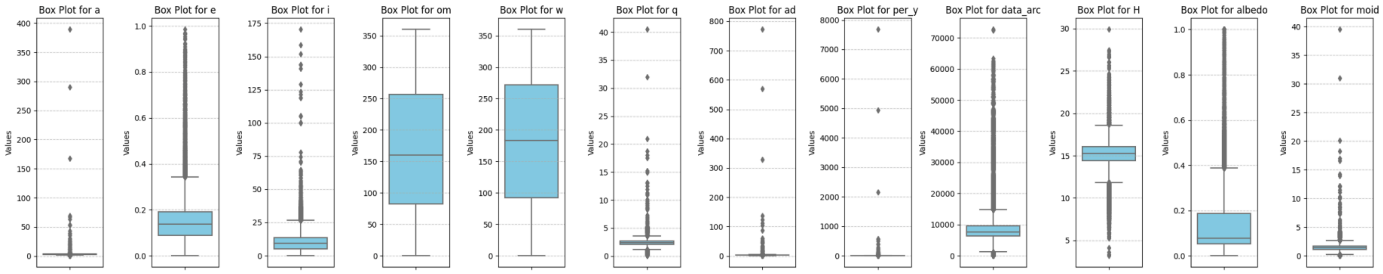

Figure 1: Box plot for all features before data cleaning

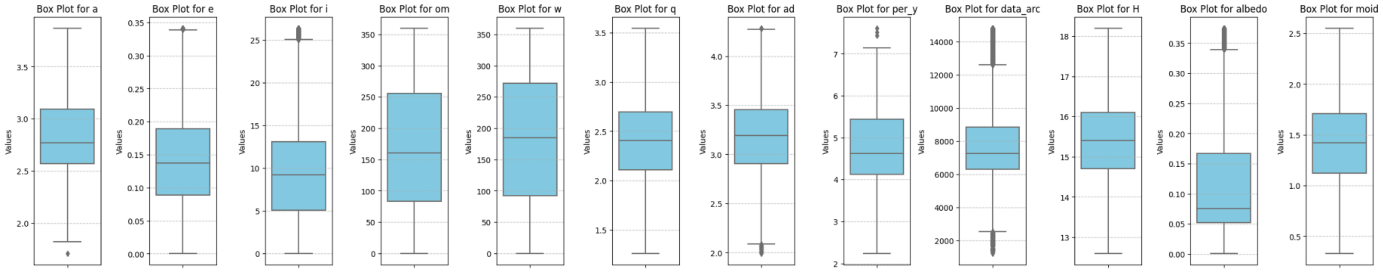

Figure 2: Box plot for all features after data cleaning

Once the data has been cleaned, we start to feature engineer the data. We first identified the ‘inter-correlation’ - correlation between the features and the target variable and ‘intra-correlation’ - correlation among the features. We then filtered the features if they lie between the threshold range of

(-0.2, 0.2). For all the features outside this range, they will be directly used in the model but for the features that lie in this range, then we combined them using PCA with suitable value for *n_components*.

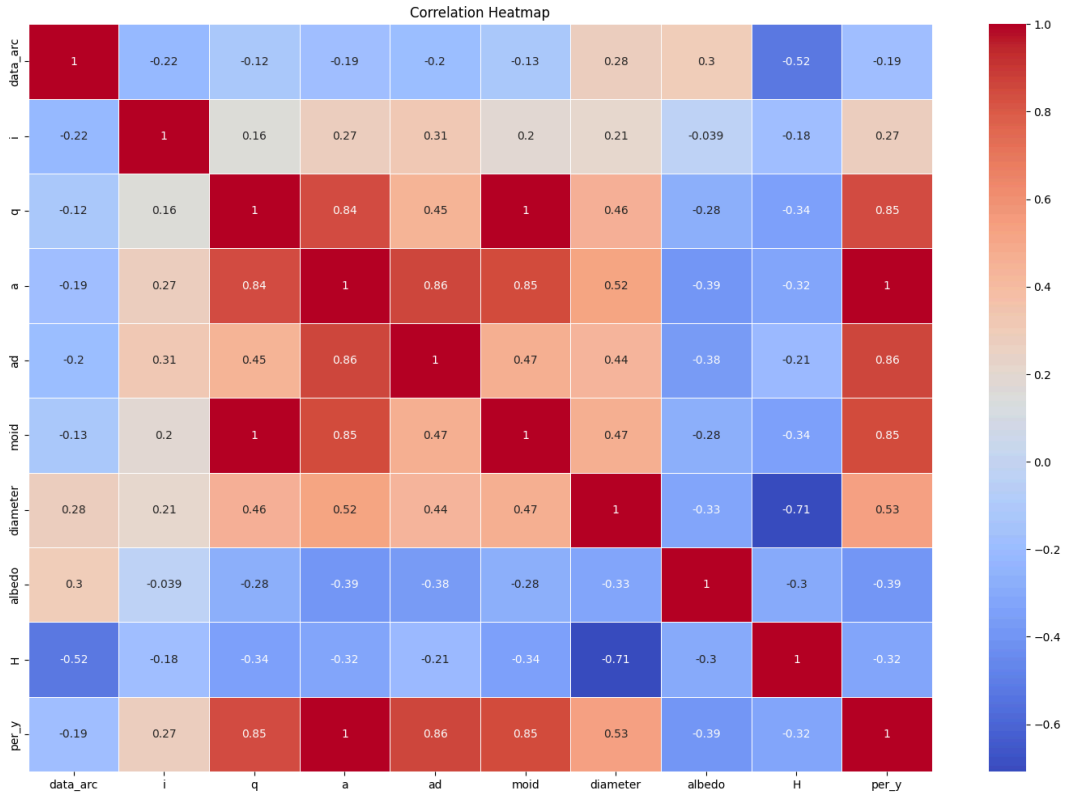

Figure 3: Correlation Map for all the numerical features

The categorical features in the data are evaluated based on pearson's coefficient, they are one-hot encoded and then used in the modeling.

2.5 Modeling

Once the data has been cleaned, feature engineered and built a pipeline to serve the models. We started to experiment across different models based on the hypothesis we formulated from our EDA. Since, most of the features are of Gaussian Distribution, using a linear model will help us a lot in our case.

As per the ***Central Limit Theorem***, once the quantity of data approaches a bigger limit all distributions approximate to gaussian distribution. Based on the above fact we scaled the data using Standard Scaler. To get a baseline reading we used ***LazyPredict's*** [3] pipeline to serve the data to most of the Machine Learning models accessible through Python.

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken
LGBMRegressor	0.96	0.96	0.46	0.41
RandomForestRegressor	0.96	0.96	0.47	73.36
XGBRegressor	0.96	0.96	0.47	1.29
ExtraTreesRegressor	0.96	0.96	0.47	22.25
SVR	0.95	0.95	0.48	282.97
GradientBoostingRegressor	0.95	0.95	0.48	28.83
BaggingRegressor	0.95	0.95	0.49	7.19
KNeighborsRegressor	0.92	0.92	0.63	4.61
DecisionTreeRegressor	0.91	0.91	0.67	1.18
PoissonRegressor	0.88	0.88	0.79	0.62
LassoLarsCV	0.84	0.84	0.90	0.23
RidgeCV	0.84	0.84	0.90	0.11
LassoCV	0.83	0.83	0.91	0.72
ElasticNetCV	0.83	0.83	0.91	0.61
OrthogonalMatchingPursuitCV	0.83	0.83	0.92	0.13
HuberRegressor	0.83	0.83	0.93	0.70
LinearSVR	0.82	0.82	0.94	2.78
LarsCV	0.82	0.82	0.95	0.21
RANSACRegressor	0.78	0.78	1.06	0.17
TweedieRegressor	0.67	0.67	1.29	0.54
AdaBoostRegressor	0.67	0.67	1.29	8.64
PassiveAggressiveRegressor	0.60	0.60	1.41	0.20
SGDRegressor	-9729206553834004.00	-9718871327171312.00	220380594.87	0.19

Figure 4: LazyPredict

Based on the above experimentation we started to ***Fine-Tune*** the top performing models from lazypredict. We chose ***DecisionTreeRegressor***, ***XGBoost , RandomForestRegressor, LinearRegressor and LightGBM***

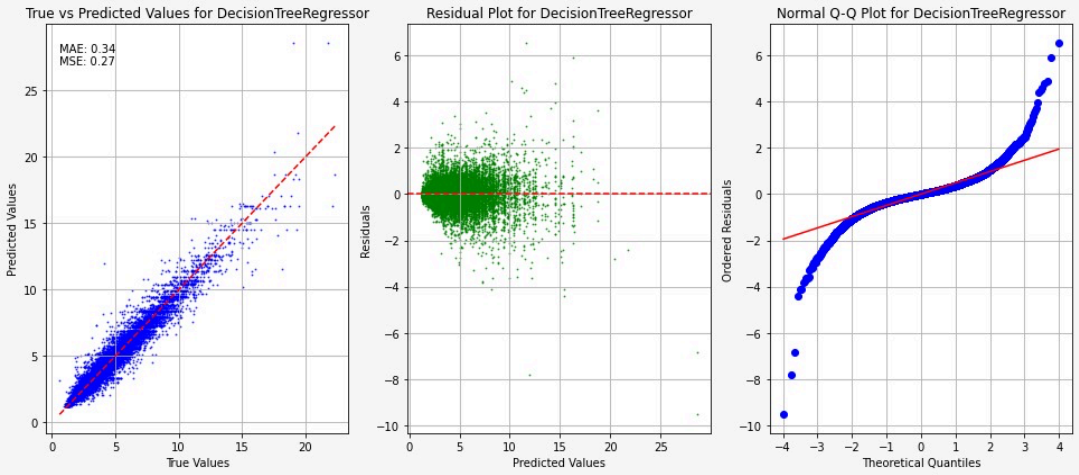

Figure 5 :(a) Scatter Plot (b) Decision Tree (c) QQ plot for Decision Tree Regressor

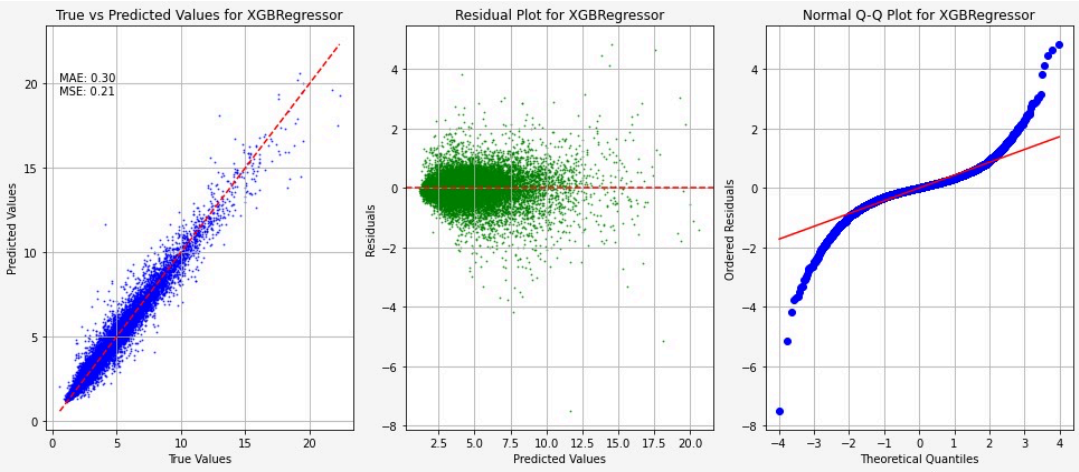

Figure 6:(a) Scatter Plot (b) Decision Tree (c) QQ plot for XGBoost Regressor

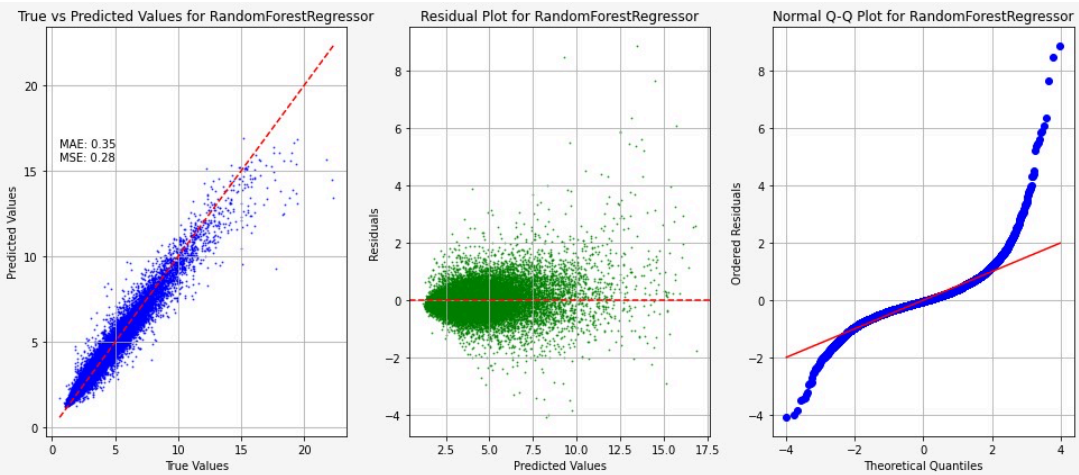

Figure 7 :(a) Scatter Plot (b) Decision Tree (c) QQ plot for RandomForest Regressor

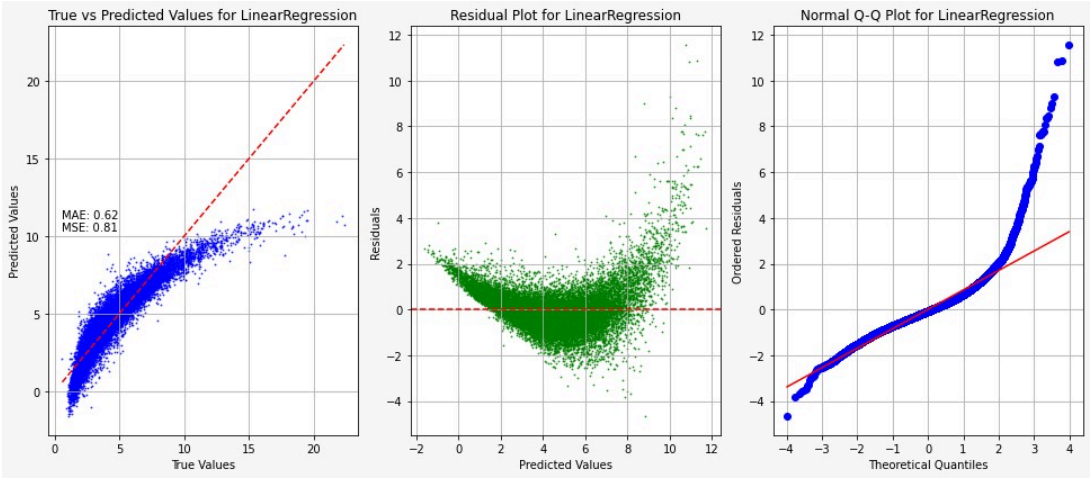

Figure 8 :(a) Scatter Plot (b) Decision Tree (c) QQ plot for Linear Regression

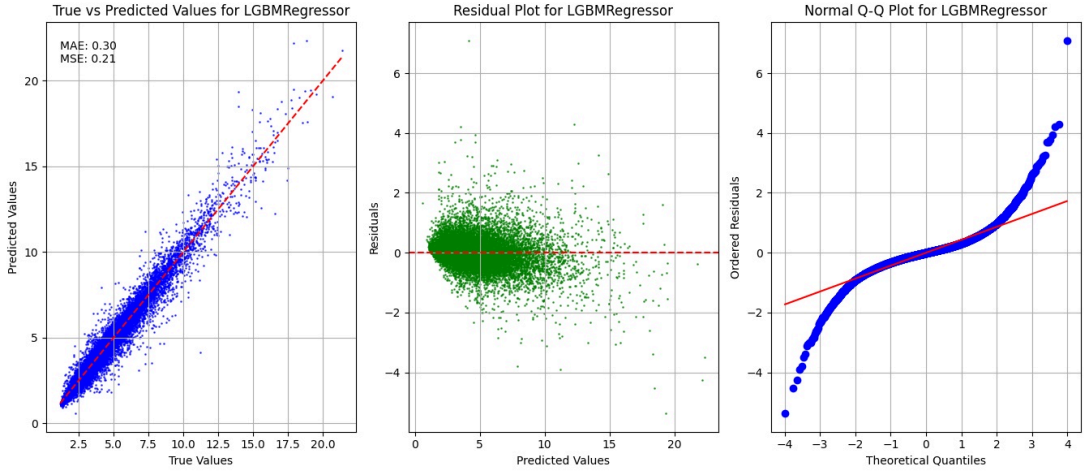

Figure 9 :(a) Scatter Plot (b) Decision Tree (c) QQ plot for LightGBM

3. Results

From the JPL database before using the data for the modeling we have taken out the test set. For that test set we have applied all the models and we got the following metrics.

Mean Absolute Error (MAE): MAE measures the average absolute difference between the predicted values and the actual values.

Mean Squared Error (MSE): MSE measures the average squared difference between the predicted values and the actual values.

R-squared: R-squared is a statistical measure that represents the proportion of the variance in the dependent variable that can be explained by the independent variables.

Adj R-squared: Adjusted R-squared is a penalized R-squared that adjusts for model complexity by considering the number of features used.

Regression Model	MAE	MSE	R-squared	Adj R-squared
Linear	0.61	0.81	0.83	0.83
Decision Tree	0.33	0.26	0.94	0.94
Random Forest	0.35	0.28	0.94	0.94
XGBoost	0.29	0.20	0.95	0.95
LightGBM	0.30	0.20	0.96	0.96

Table 1: Evaluation Metrics

Our analysis shows that tree-based models (Decision Tree, Random Forest, XGBoost, and LightGBM) outperform Linear Regression based on both MAE and MSE. This suggests they capture the underlying patterns in the data more effectively, leading to lower prediction errors. Among the tree-based models, LightGBM shines. It boasts the lowest errors (MAE and MSE) and the highest explanatory power (R-squared and Adjusted R-squared), suggesting it captures the data's complexity slightly better than its counterparts.
In conclusion, LightGBM achieves the lowest Mean Absolute Error (MAE) of 0.30 and the lowest Mean Squared Error (MSE) of 0.20. This indicates that, on average, LightGBM's predictions deviate less from the actual asteroid diameters compared to other models.

4. Discussion

The results of this asteroid project hold significant meaning and impact in several key areas. By improving the accuracy of asteroid diameter predictions, the project enhances our understanding of asteroid dynamics and aids in mitigating potential threats to Earth. This knowledge is invaluable for researchers, astronomers, and space agencies, enabling them to make better-informed decisions regarding asteroid detection, threat assessment, and mitigation strategies.

The findings from this project can benefit a wide range of stakeholders. For example, space agencies can utilize the predictive models and data analysis techniques developed in this project to enhance early warning systems for potential asteroid impacts. Researchers in planetary science can leverage the insights gained to further explore the dynamics of asteroids and their impact on the solar system's evolution. Additionally, the general public can benefit from improved understanding and awareness of asteroid threats, empowering individuals and communities to take proactive measures for disaster preparedness.

The results of this project can be used to make better-informed decisions in various ways. For instance, the predictive models can aid in prioritizing asteroid detection efforts, focusing resources on asteroids with the highest probability of impacting Earth. Furthermore, the insights gained from analyzing the factors influencing asteroid diameter predictions can inform the development of more accurate and robust predictive models, leading to more effective threat mitigation strategies.

In future work, several aspects of the project could be improved. For example, incorporating additional data such as absolute magnitude (H) and orbital parameters and additional datasets could further refine predictive models, improve accuracy. Additionally, exploring advanced machine learning algorithms and

feature engineering techniques may enhance the predictive capabilities of the models. Moreover, expanding the scope of the project to include a broader range of asteroids and incorporating real-time data could provide more comprehensive insights into asteroid dynamics and behavior.

5. Statement of contributions

This project is a testament to the teamwork and shared effort of all authors. Each individual played a crucial role in achieving the project's objectives.

Aditya Shanmugham	Data Preprocessing, XGBoost, Exploratory Data Analysis, Data Sourcing
Jonathan Paul Dande	LightGBM, Linear Regression, Exploratory Data Analysis
Srinidhi Sunkara	Feature Selection, Random Forest and Decision Tree Regression Results, EDA

6. References

1. Dataset - https://www.kaggle.com/datasets/basu369victor/prediction-of-asteroid-diameter
2. https://academic.oup.com/mnras/article/499/3/4570/5918397
3. https://lazypredict.readthedocs.io/en/latest/

7. Appendix

1. Codebase - https://github.com/Srinidhi-Sunkara/Asteroid_Diameter_Detection
2. Life Cycle -

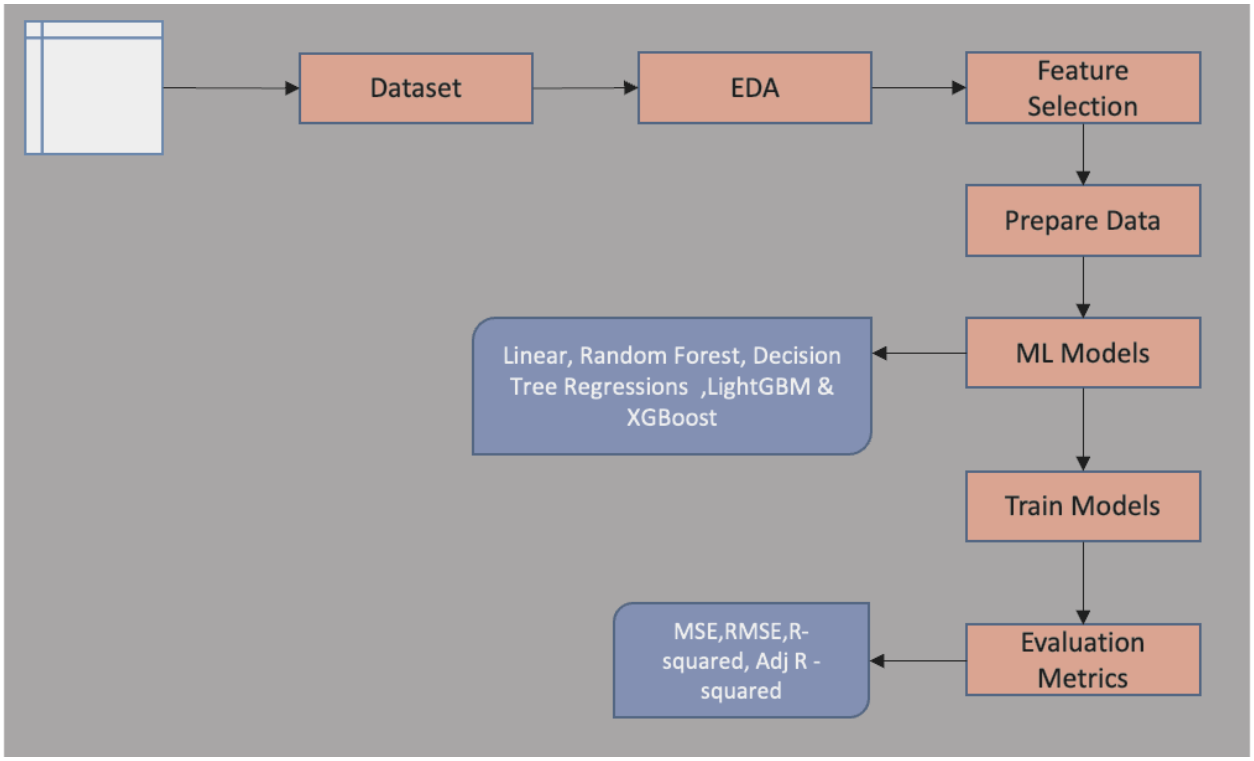

Figure 10: Life Cycle