Health Insurance

By:
Kelsey Woods &
Aayushi Verma

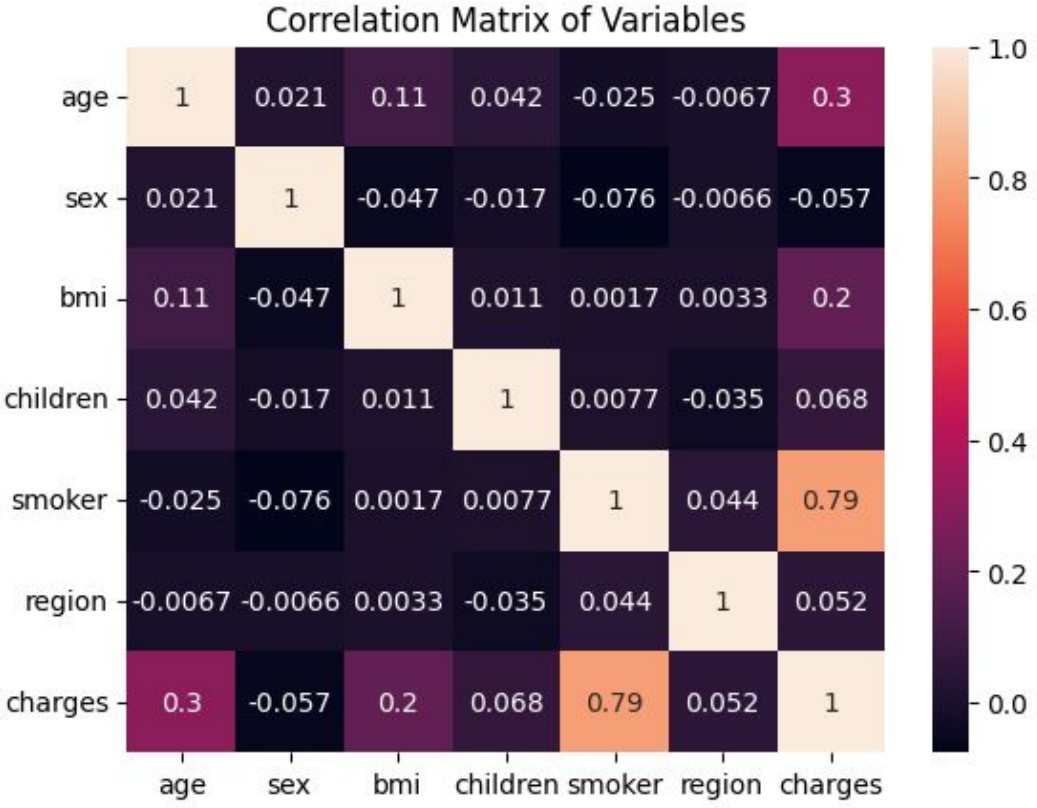
Problem Statement & Dataset

- We are using a dataset titled Medical Cost Personal Dataset by Kaggle
- It lists the cost per family and its possibly associated factors, such as age and BMI
- We are looking to see if we can find the associated causes of a higher insurance plan

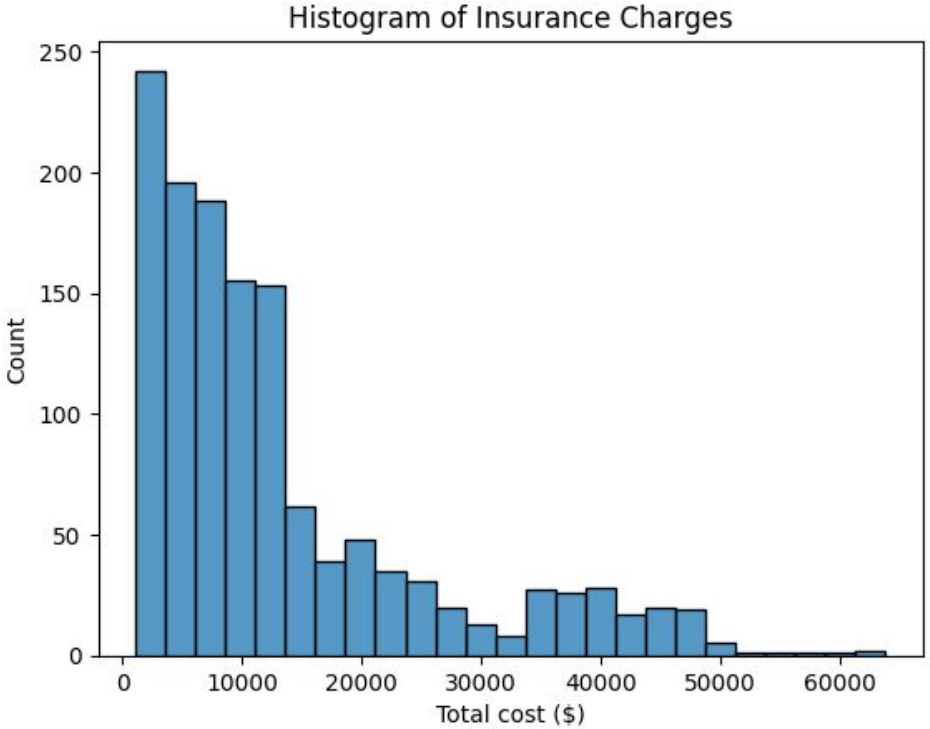
Data Cleaning

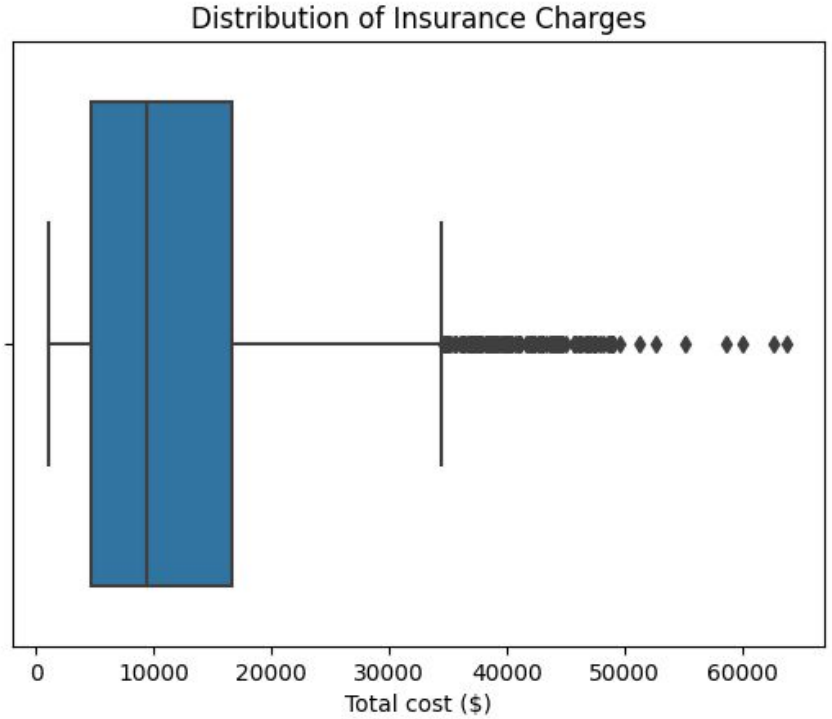
- Removed NaN, null, missing values
- Removed duplicate values
- Imputed NA values with mean value of column
- Performed label encoding to change from categorical data to numeric data type

Exploratory Data Analysis

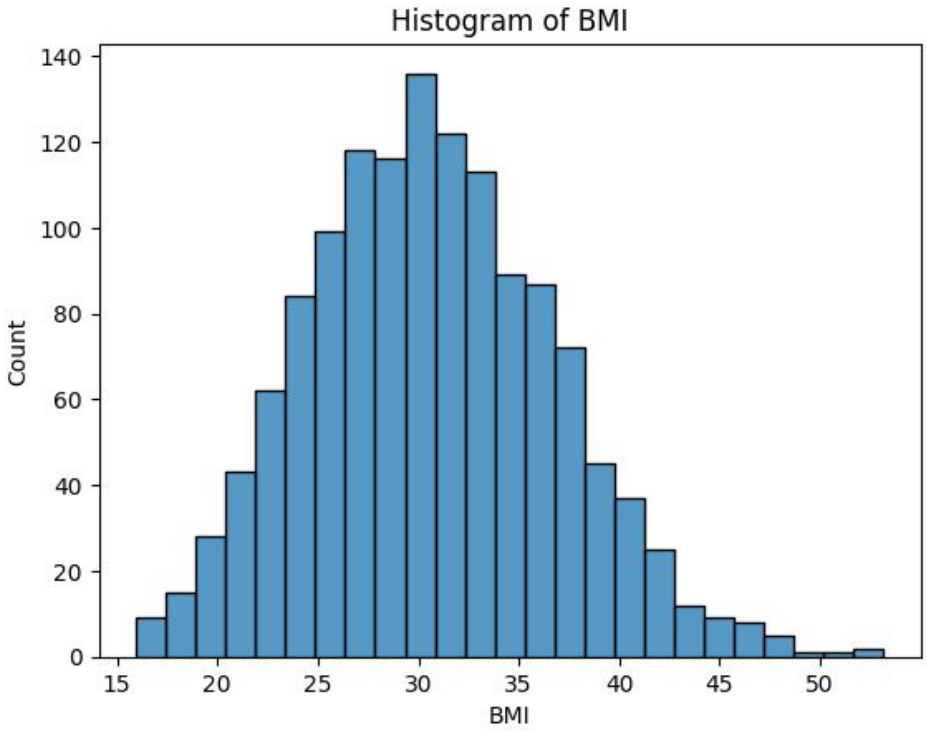


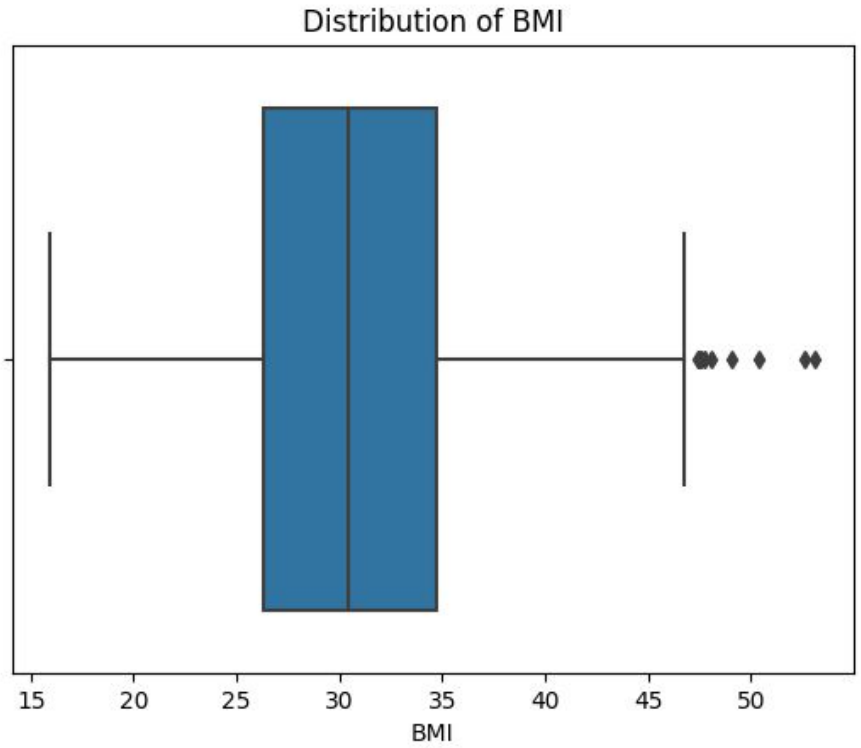
Exploratory Data Analysis



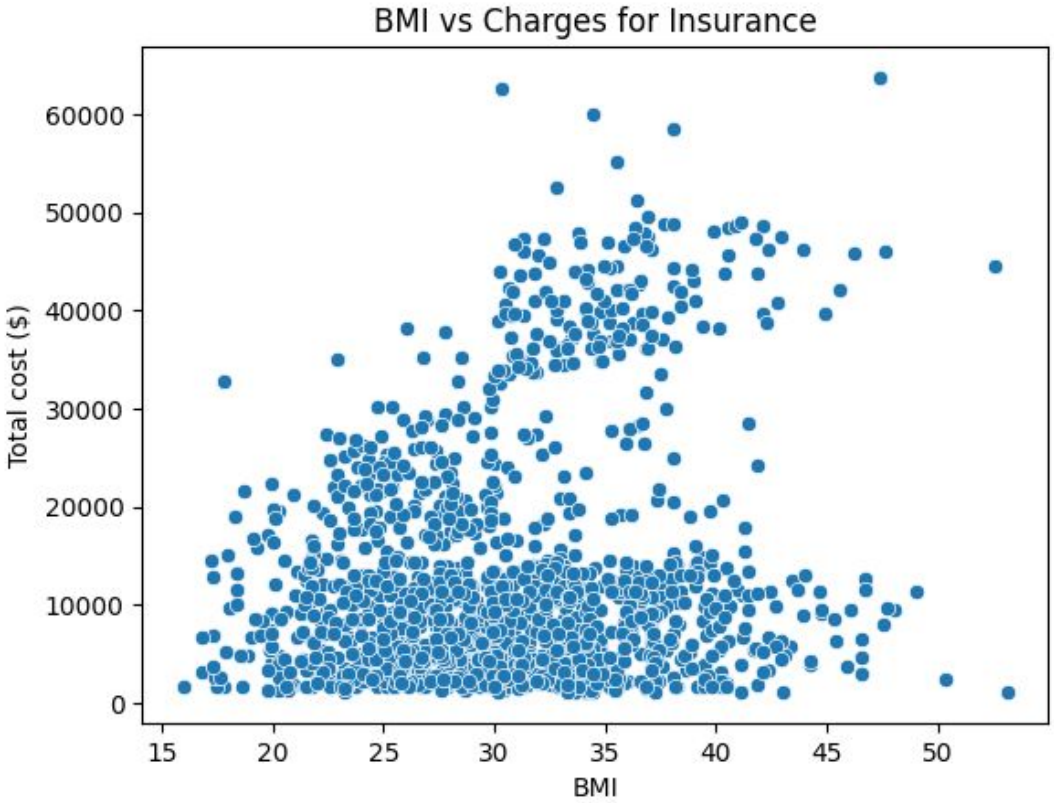


Exploratory Data Analysis





Exploratory Data Analysis



Machine Learning Algorithms

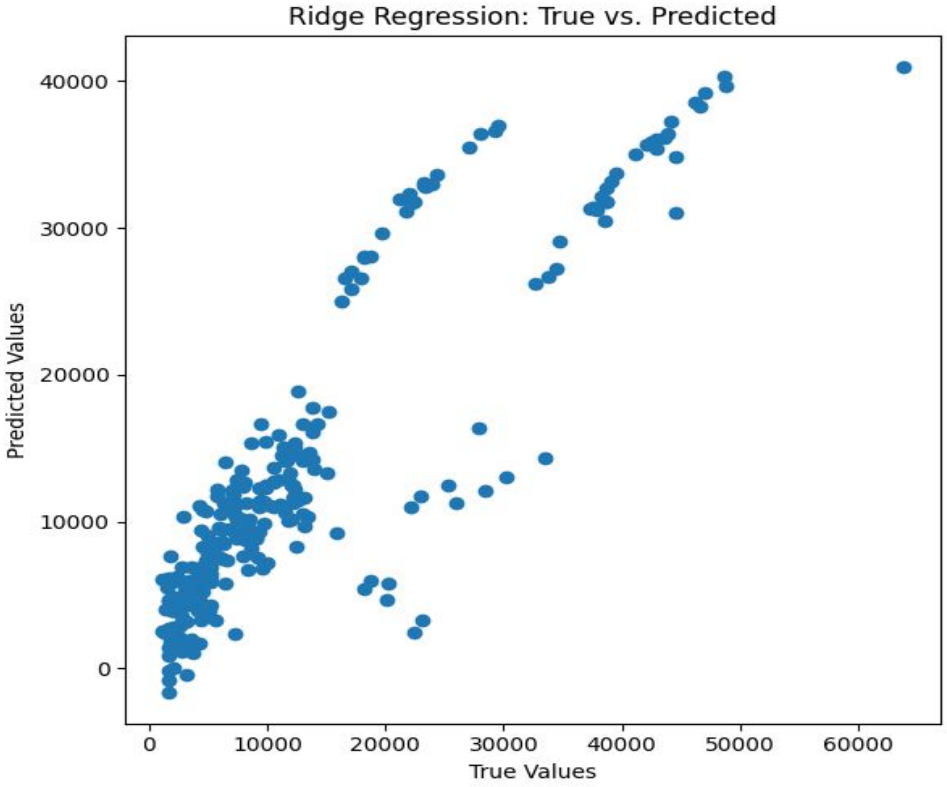
Linear Regression
Random Forest
SVM
Lasso Regression
Decision Tree

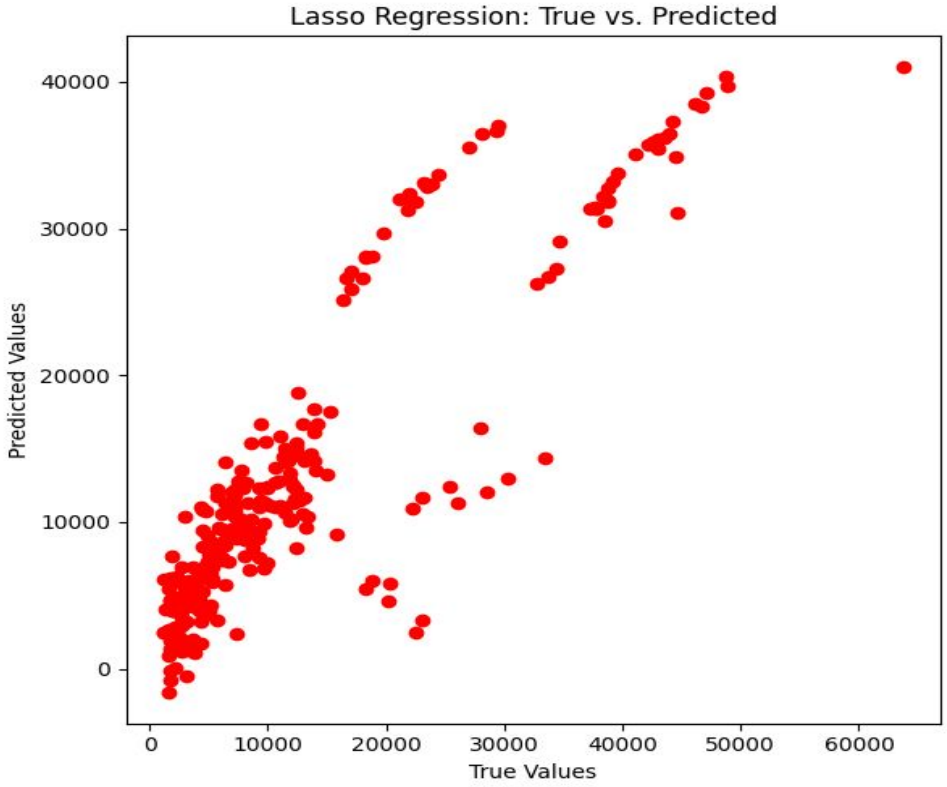
```
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.2,
random_state=42)
```

Optimization

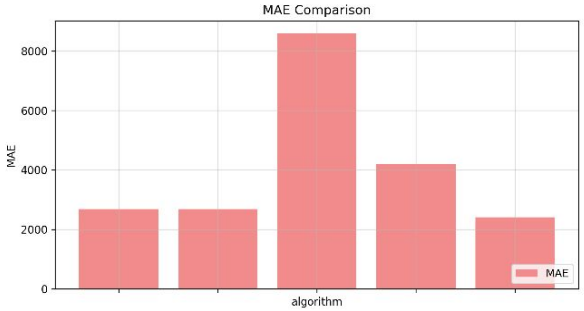
	ml_algorithm	mae	mse	rmse	r2	adjusted_r2
0	Decision Tree Regressor	2684.822145	3.535392e+07	5945.916022	0.772276	0.771249
1	Linear Regression	2684.822145	3.392761e+07	5824.741112	0.781463	0.780478
2	Support Vector Machine	8592.435460	1.664724e+08	12902.418789	-0.072295	-0.077129
3	Lasso Regression	4207.104528	3.393092e+07	5825.025724	0.781442	0.780456
4	Gradient Boosting Regressor	2416.156153	1.897998e+07	4356.601777	0.877745	0.877194

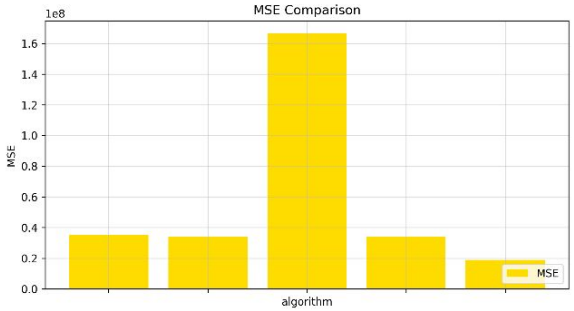
Hyperparameter Tuning

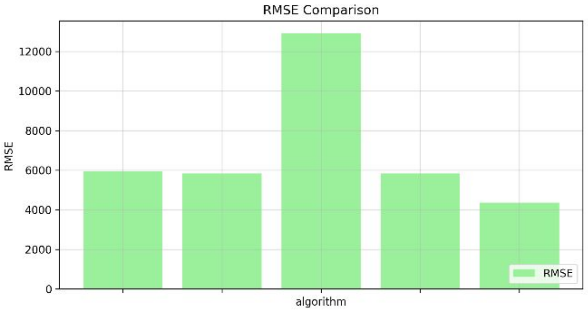


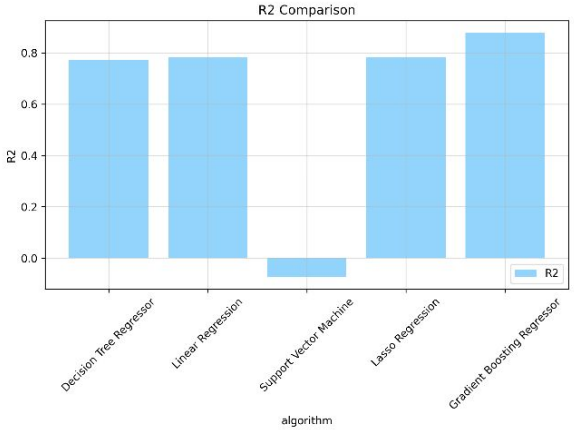


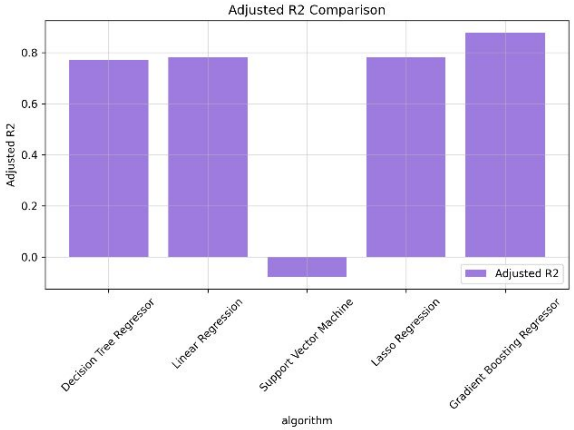
Model Evaluation & Comparison

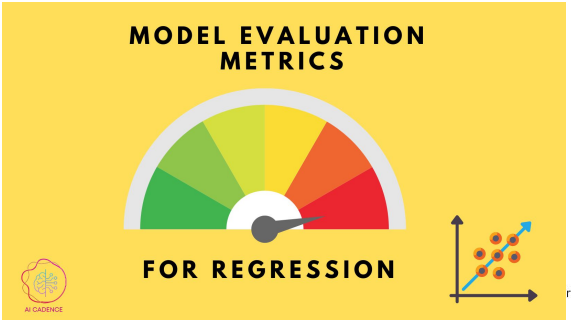




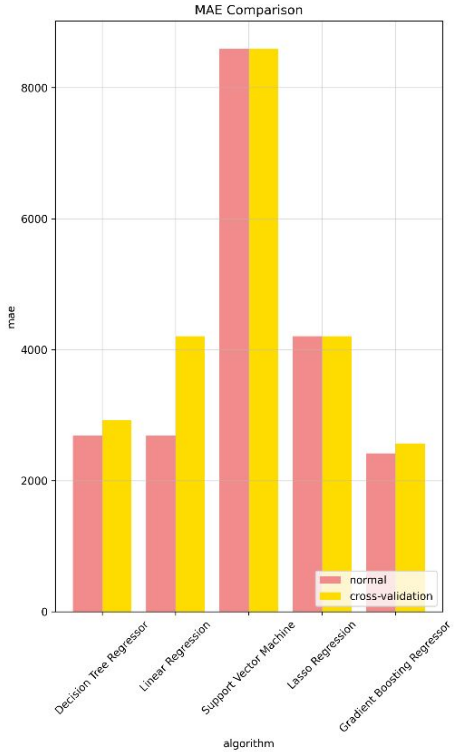


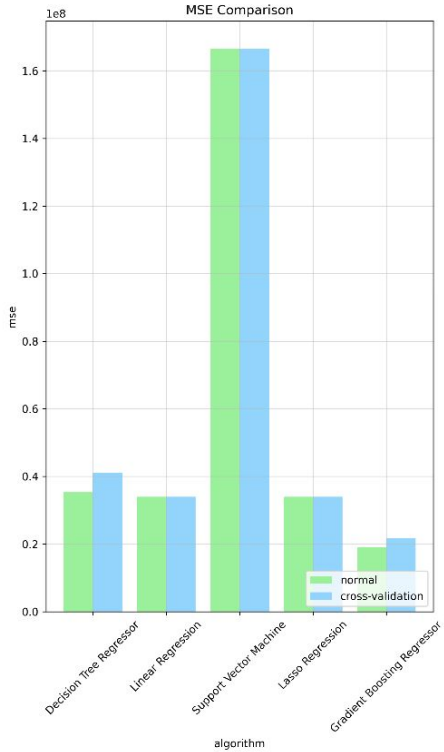


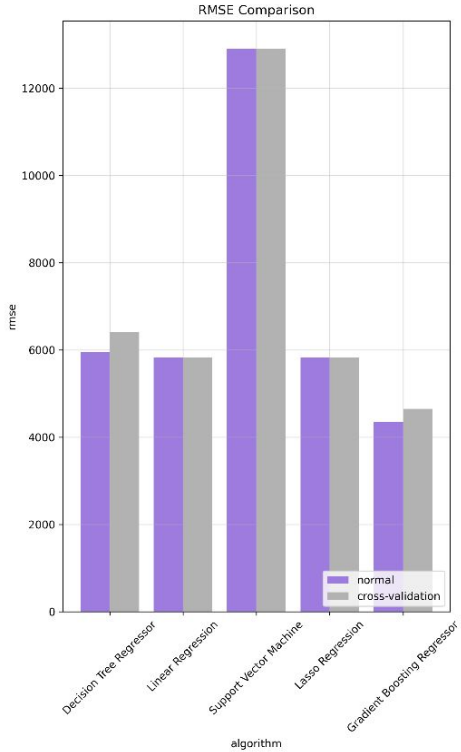


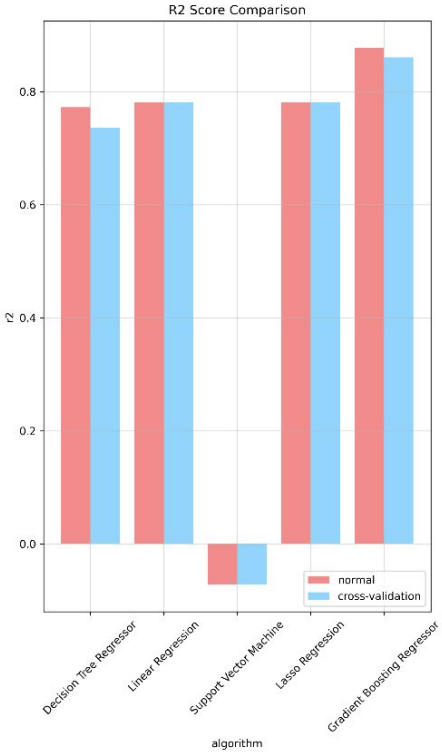


Model Evaluation & Comparison

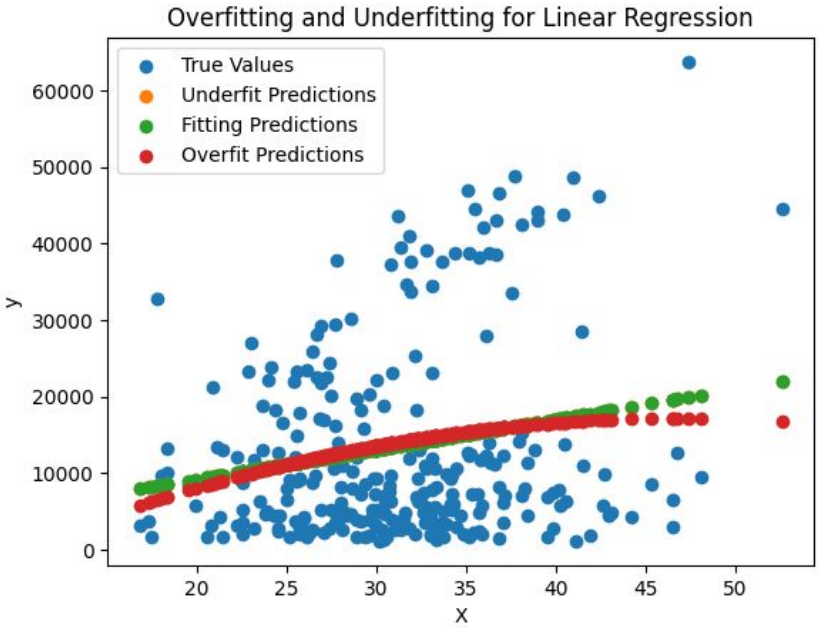


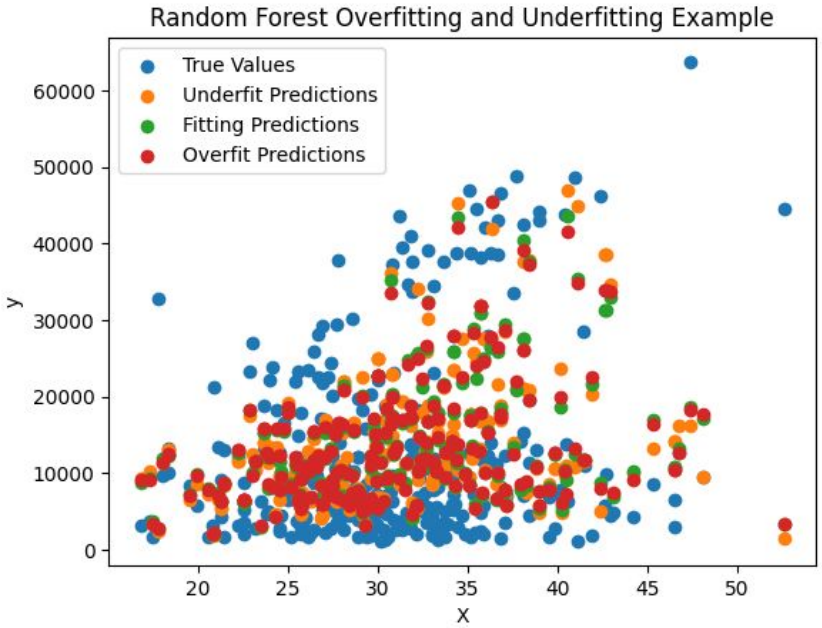




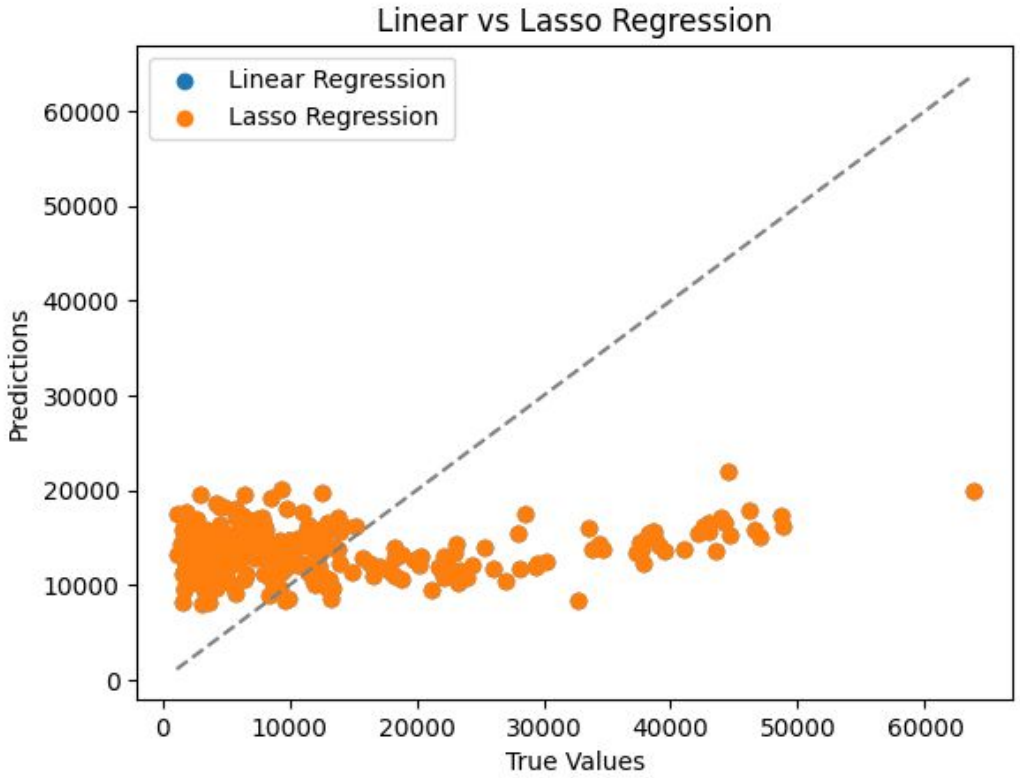


Model Evaluation & Comparison





Model Evaluation & Comparison



Discussion & Conclusion

- Gradient boosting was the best model, SVM was the worst.
- The hyperparameter tuning and gradient boosting significantly helped improve the performance of the model because…..

Hyperparameter Tuning	Gradient Boosting
Lessens the effect of overfitting/underfitting	Combines the strengths of multiple weak learners by using other models to provide strength in that area
It can lead to improved accuracy, precision, recall, and other performance metrics	Captures complex patterns (including non-linear ones)
It can reduce training time and need for resources	It helps with feature selection and managing unbalanced data