**SUMMARY**

This is the bird’s-eye view for the Project flow.

1. Data Processing
2. Feature 10: Logarithm of the Dependent variable is taken to approximate the normal distribution
3. Feature 5 – 8:  
   For these categorical variables, first new groups been defined and test with ANOVA to check whether these groups are statistically different from each other
4. Feature 1 – 4 & Feature 9 (For independent continuous variable):

Min-Max scaling is done to standardize the data before modelling

1. Dummy variables are created for Feature 5 – 8 with reduced statistically different bins
2. Model Evaluation

Following Models were run with the given metrics to check best score

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  | R^2 Score | | MAE |
| Model | Add On | Hyper parameter Tuning | Train | Test | Train |
| Linear Regression | Simple | NA | 0.77 | “-ve” | 0.2632 |
| Ridge | Grid Search | 0.795 | 0.793 | 0.2638 |
| Lasso | Grid Search | 0.8 | 0.79 | 0.3408 |
| KNN | 1 Neighbor | NA | 0.6279 |  |  |
| 2 Neighbors | NA | 0.7101 |  |  |
| 3 Neighbors | NA | 0.7375 |  |  |
| 4 Neighbors | NA | 0.7373 |  |  |
| Polynomial | Degree 2 | NA | 0.868 | "-ve" | 0.2088 |
| Ridge | Grid Search | 0.83 | 0.81 | 0.22 |
| Lasso | Grid Search | 0.768 | 0.804 | 0.792 |
| SVM | Linear Kernel | Grid Search | 0.793 | 0.792 | 0.2612 |
| RBF Kernel | Grid Search | 0.791 | 0.785 | 0.2464 |
| RBF Kernel | Genetic Algorithm | 0.423 | 0.425 | NA |

1. Best Model came out to be Polynomial regression with Ridge
2. Following files are attached as submission -   
   Python Code, Word Summary Document and Predictions for verification for each IDs