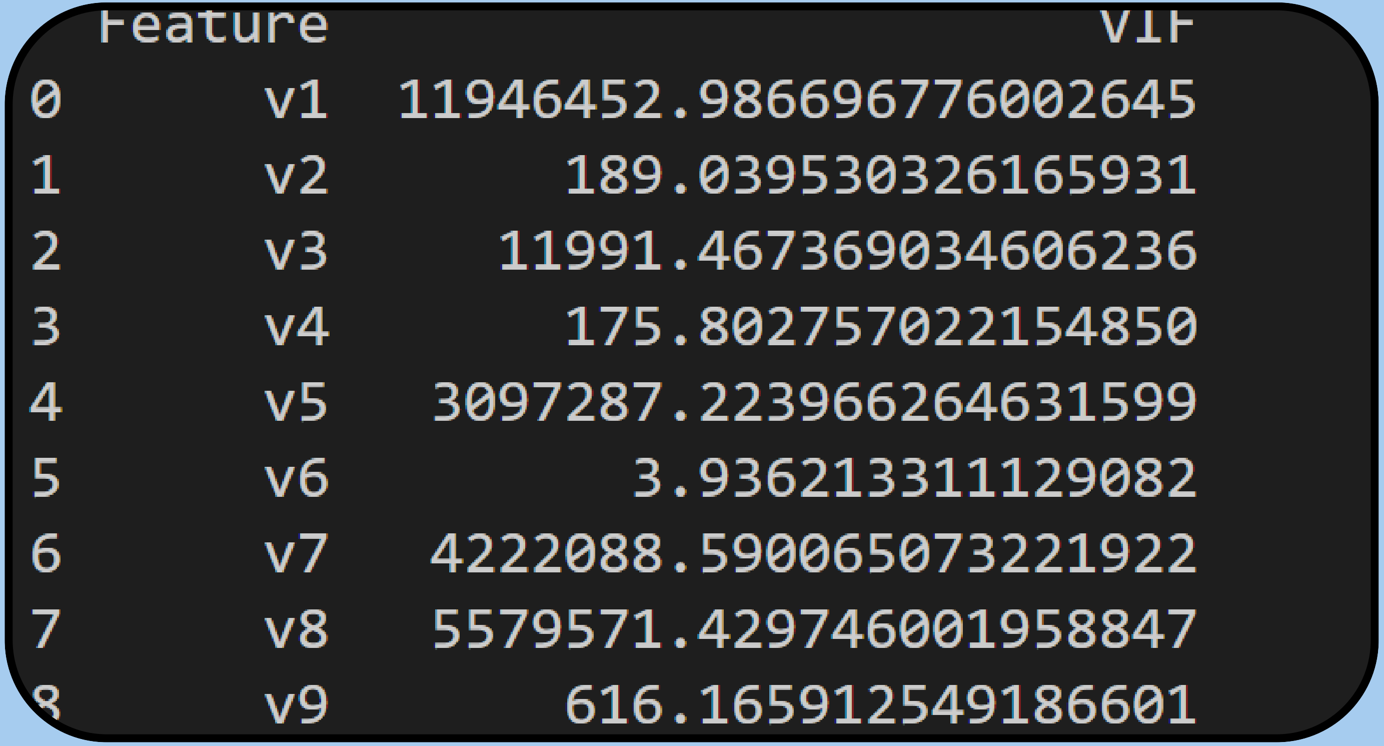
STATISTICAL LEARNING & DATA MINING

Parametric and Non-Parametric : Linear Regression and KNN

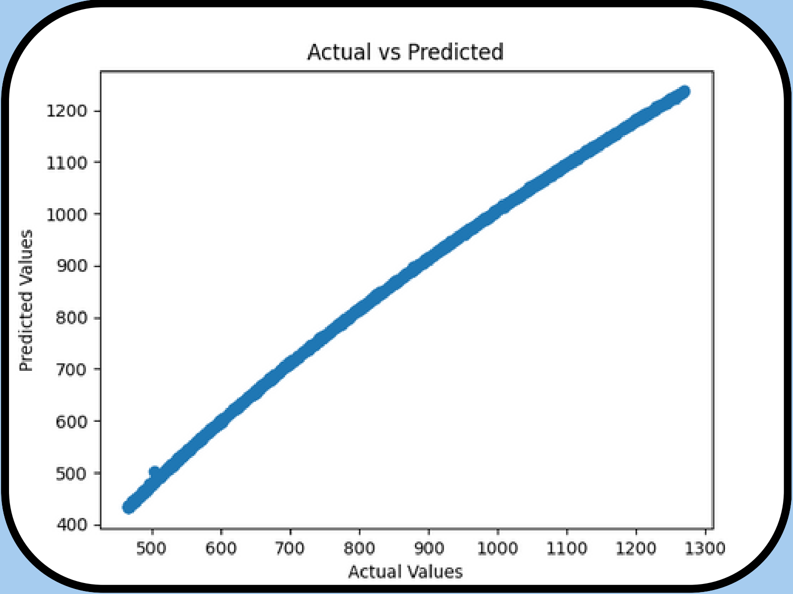
MAIA

By Madinabonu Akramova

DATA PREPROCESSING



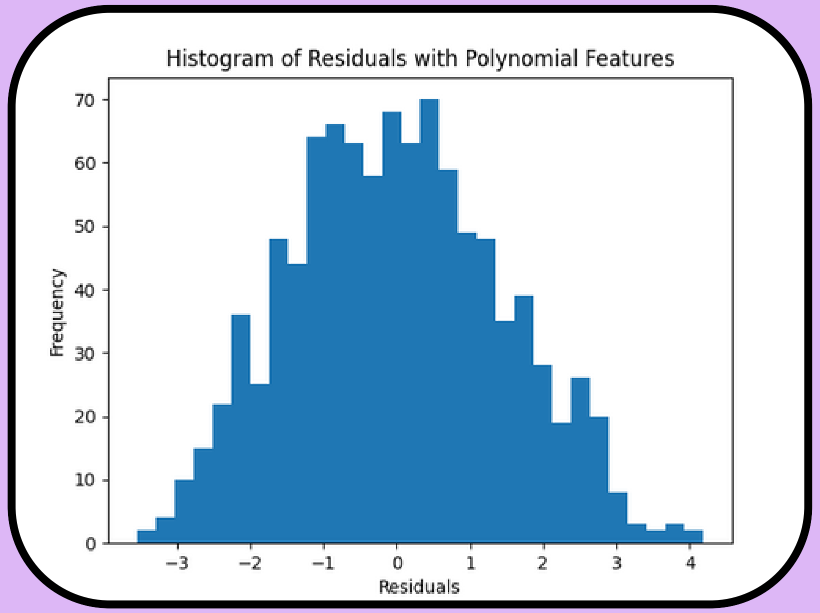
```
  Feature                      VIF
0      v1   11946452.986696776002645
1      v2        189.039530326165931
2      v3      11991.467369034606236
3      v4        175.802757022154850
4      v5    3097287.223966264631599
5      v6          3.936213311129082
6      v7    4222088.590065073221922
7      v8    5579571.429746001958847
8      v9        616.165912549186601
```

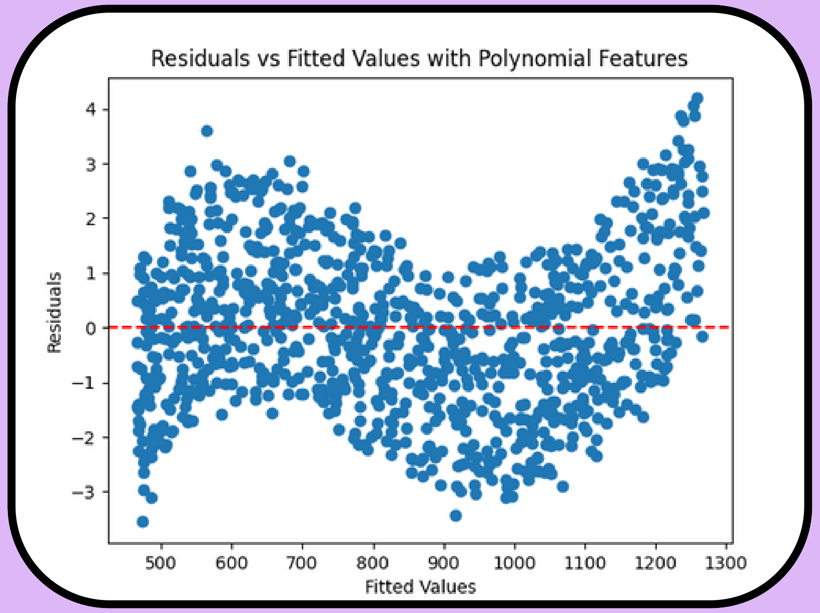


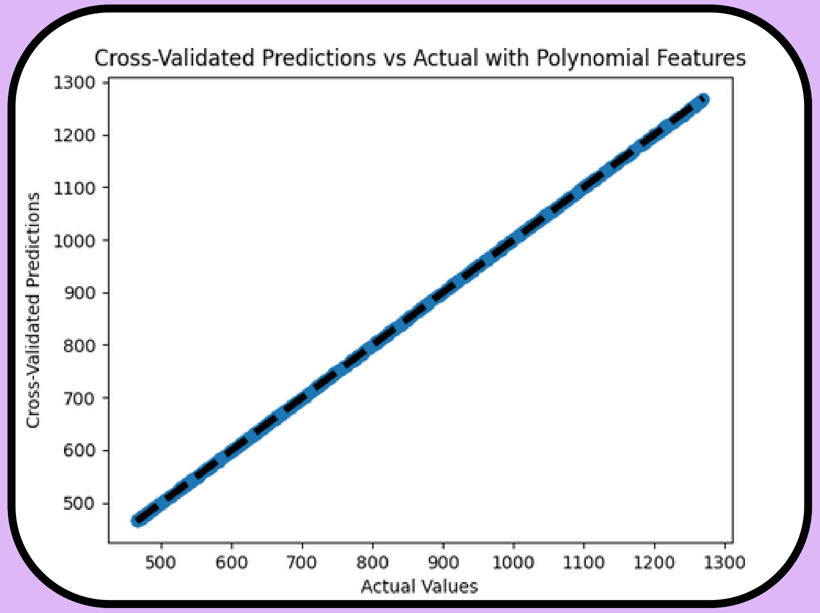
- Loading and cleaning data from index columns
- Plotting scatterplots for every value and finding that v3 shows low multicollinearity
- Identifying and removing outliers with z-score
- Applying standardization to scale features since KNN model is sensitive to data scaling
- Detecting multicollinearity using VIF (Variance Inflation Factor) and removing v1,v5,v7,v8 since their VIF was high
- Capturing non-linearity with the help of polynomial features

MODEL DEVELOPMENT & EVALUATION

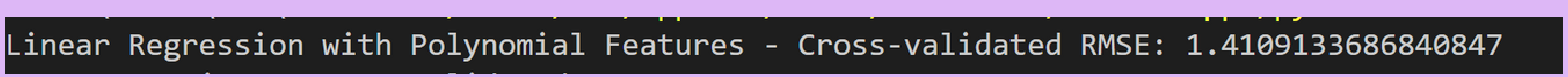
Parametric Approach


Histogram of Residuals with Polynomial Features


Residuals vs Fitted Values with Polynomial Features


Cross-Validated Predictions vs Actual with Polynomial Features

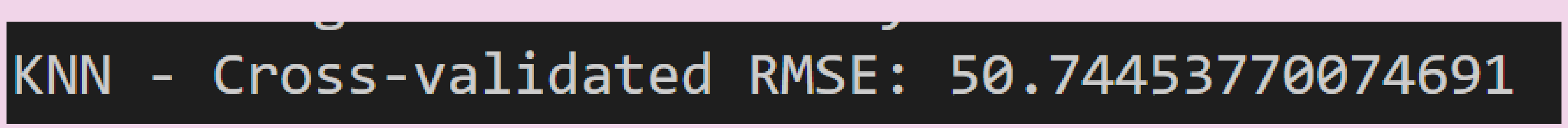
- Capturing non-linearity with the help of polynomial features because model is intended to capture complicated patterns
- Residual analysis was done for diagnostics of how well data fit and how it was able to capture complex patterns after adjustment
- RMSE calculation gave ≈ 1.41 which means it's highly accurate.



```
Linear Regression with Polynomial Features - Cross-validated RMSE: 1.4109133686840847
```

MODEL DEVELOPMENT & EVALUATION

Non-Parametric Approach



```
KNN - Cross-validated RMSE: 50.74453770074691
```

- Tuning KNN with the help of cross-validation for neighbor optimization to improve his predictive ability
- Since it was not well adjusted like Linear Regression it gave ≈ 51 which means lower accuracy but not bad

FINAL PREDICTION

- Final predictions were combined and saved
- After comparing both models it can be seen that Linear Regression did pretty good with handling non-linear patterns. KNN could have been better if it was adjusted more