Summer Project: Documentation

1. Data Cleaning:

* Data set has 6670 samples with 13 features.
* It has no missing values
* Numerical features are Acceleration [Acc], Relative velocity [DeltaV], Absolute velocity [AbsSpd], Spacing [S] features.
* Categorical features are vehicle type, leader type, category [ leader location], Acceleration gap widening.
* Divided the dataset into two dataset Acceleration gap widening and Acceleration gap narrowing.
* Applied the one-hot-encoding on categorical nominal features.
* AccGapWid has 3769 samples and AccGapNar as 2901 samples and 11 features each.
* ['Acc', 'AbsSpd', 'S', 'DeltaV', 'Type-Car', 'Type-TW', 'SLType-Car', ‘SLType -TW',

'Category-Left', 'Category-NoLead', 'Category-Right'] -> Features.

* Removed unreasonable combinations of the above features:

calculated the additional feature Collision time -> S/DeltaV.

If Collision time is less than 0.9,

if (abs(DeltaV) < 0.6 & abs(Acc > 2.27)),

if (abs(DeltaV) > 3.06 & abs(Acc <0.44))

are considered as unreasonable combinations and removed from the dataset.

There are about 230, 207 such samples for widening and narrowing respectively.

Final dataset has 3539, 2694 samples for widening and narrowing respectively.

* Binned data:
  + Implemented binning on widening and narrowing after applying all the data cleaning techniques.
  + Bins have been created for AbsSpd feature with bin size of 0.1 m/s
  + Then calculated the mean values for each bin.
  + Now the binned dataset has 144, 242 samples for widening and narrowing respectively.
  + The idea behind using bins is to reduce the noise in the dataset.

1. Model Building:

* Building Machine learning models to predict Acc.
* Built four different ML models on both binned and unbinned data for both datasets widening and narrowing.
* Applied RobustScaler preprocessing technique on all the independent features.
* Spitted the dataset into 30% test and 70% training data

1. Machine learning Models on Binned Data:

* Widening dataset

Model performance on Training the test data.

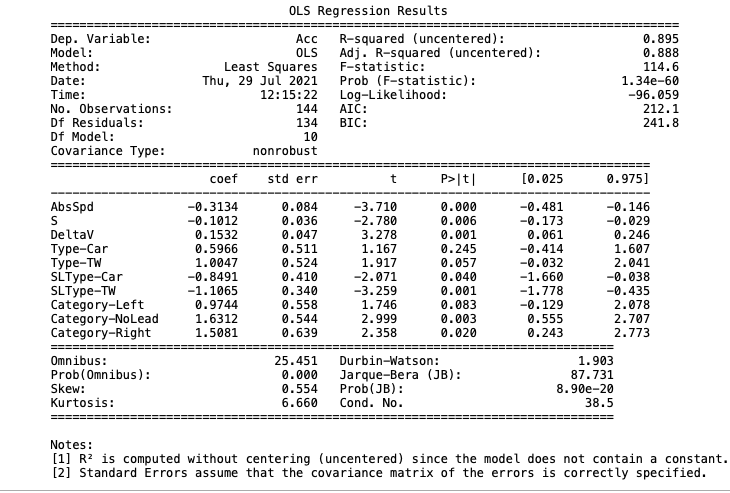
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Train\_  R2score | Test\_  R2score | Train\_  RMSE | Test\_  RMSE | Train\_  MAE | Test\_  MAE |
| XGBR | 0.389 | 0.476 | 0.487 | 0.362 | 0.294 | 0.262 |
| RF | 0.256 | 0.329 | 0.538 | 0.409 | 0.338 | 0.286 |
| SVR | 0.251 | -0.014 | 0.54 | 0.503 | 0.328 | 0.335 |
| LR | 0.35 | 0.209 | 0.503 | 0.444 | 0.338 | 0.312 |

Feature Importance:

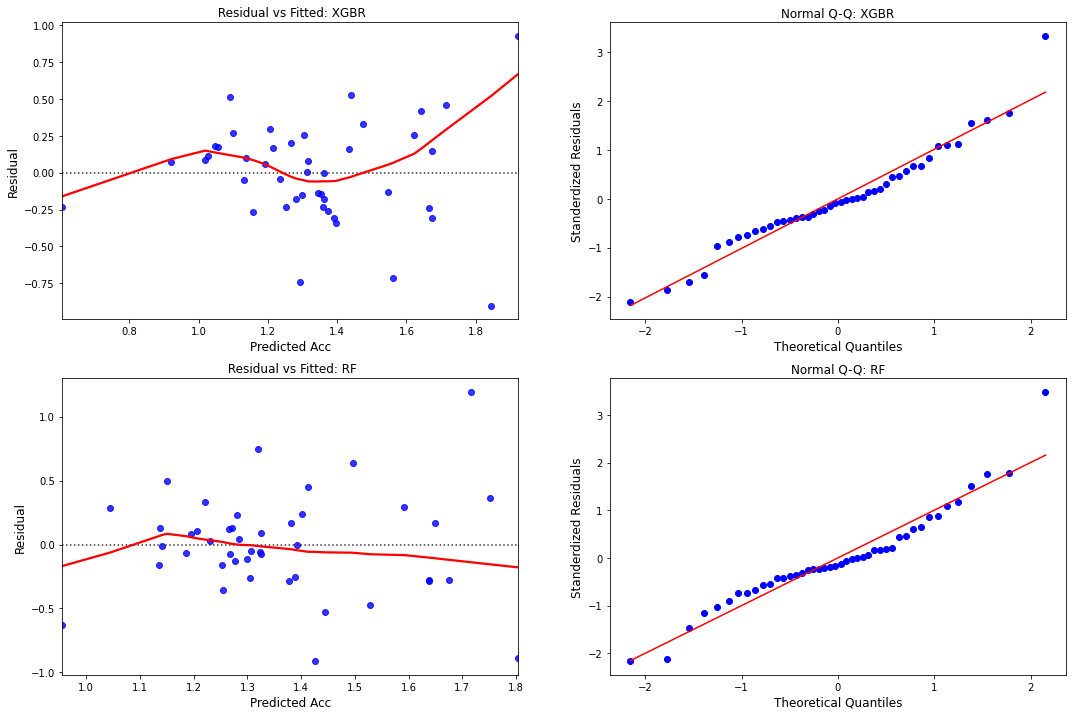
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | XGBR | RF | SVR | LR |
| AbsSpd | 0.177 | 0.378 | -0.043 | -0.221 |
| S | 0.115 | 0.106 | 0.006 | -0.108 |
| DeltaV | 0.115 | 0.183 | 0.111 | 0.126 |
| Type-Car | 0.106 | 0.064 | -0.118 | 0.027 |
| Type-TW | 0.099 | 0.045 | -0.123 | 0.113 |
| SLType-Car | 0.063 | 0.016 | -0.302 | -0.296 |
| SLType-TW | 0.071 | 0.026 | -0.389 | -0.399 |
| Category-Left | 0.11 | 0.073 | 0.158 | 0.18 |
| Category-NoLead | 0.085 | 0.056 | 0.231 | 0.375 |
| Category-Right | 0.06 | 0.053 | 0.418 | 0.497 |

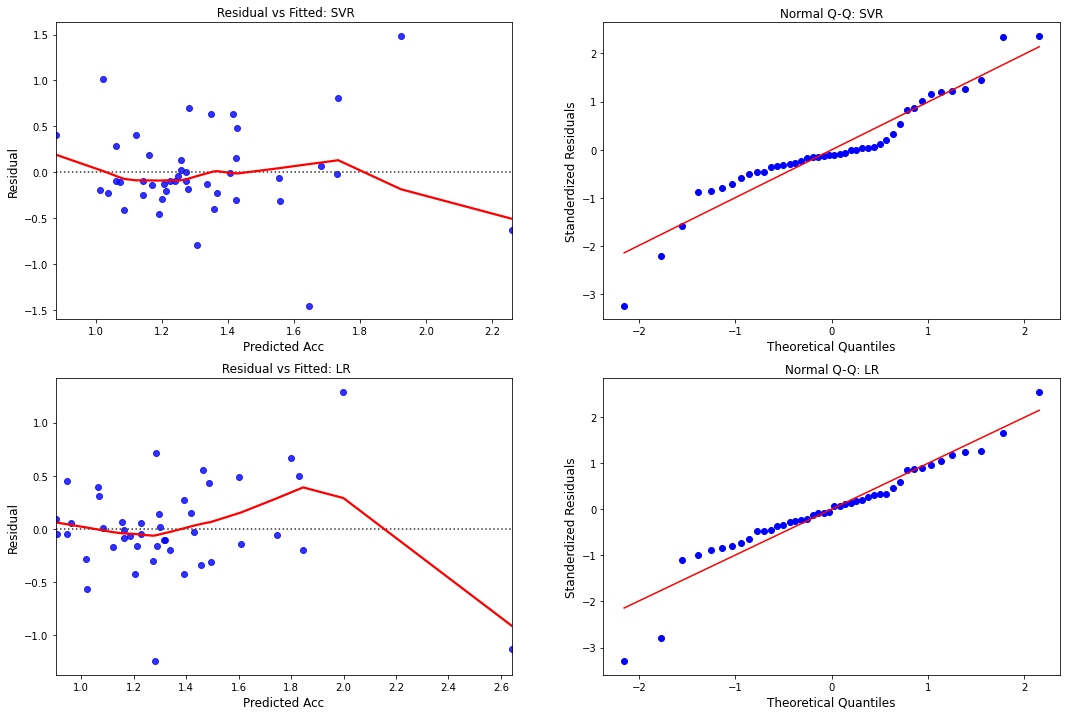
\*\* for SVR and LR option to calculate feature importance is not available, instead coefficients have been displayed.

ANOVA Table:



Residual Plots:





* Narrowing dataset

Model performance on training and test data:

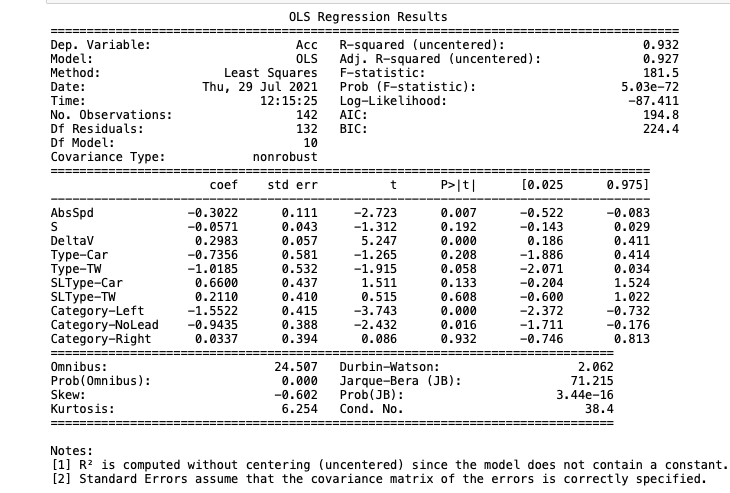
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Train\_  R2score | Test\_  R2score | Train\_  RMSE | Test\_  RMSE | Train\_  MAE | Test\_  MAE |
| XGBR | 0.735 | 0.647 | 0.395 | 0.516 | 0.254 | 0.353 |
| RF | 0.648 | 0.648 | 0.455 | 0.515 | 0.287 | 0.325 |
| SVR | 0.65 | 0.692 | 0.454 | 0.482 | 0.287 | 0.333 |
| LR | 0.675 | 0.654 | 0.437 | 0.511 | 0.294 | 0.376 |

Feature Importance:

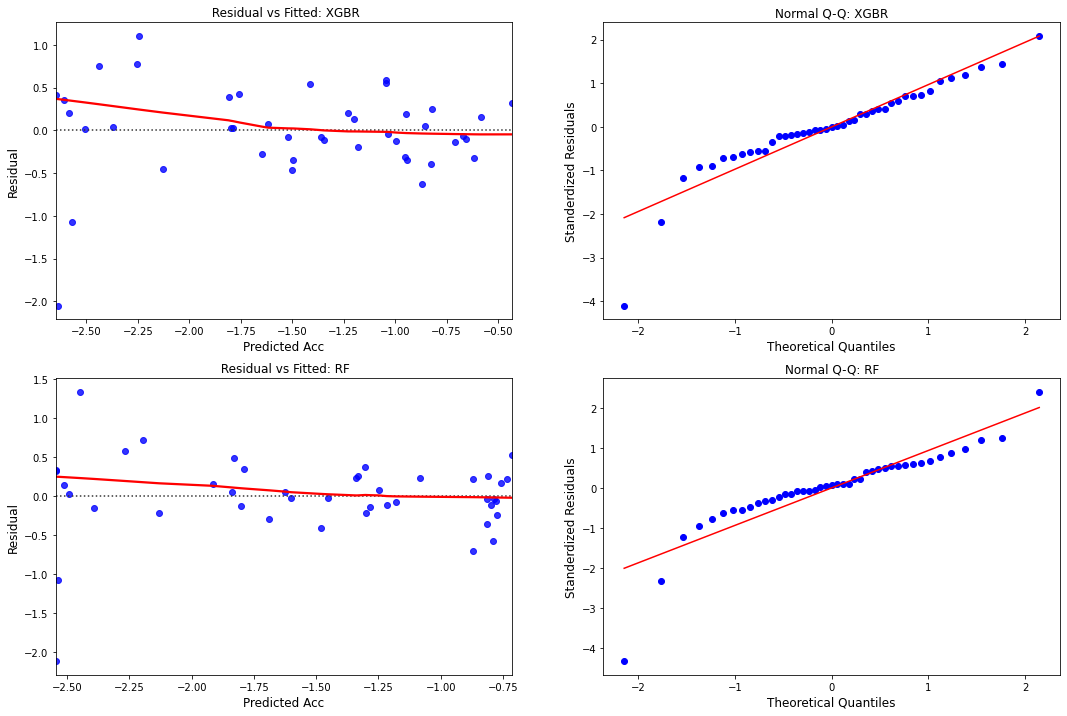
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | XGBR | RF | SVR | LR |
| AbsSpd | 0.354 | 0.512 | -0.384 | -0.296 |
| S | 0.068 | 0.018 | -0.039 | -0.088 |
| DeltaV | 0.207 | 0.390 | 0.293 | 0.226 |
| Type-Car | 0.05 | 0.005 | -0.169 | -0.324 |
| Type-TW | 0.048 | 0.000 | -0.234 | -0.398 |
| SLType-Car | 0.032 | 0.001 | 0.052 | 0.107 |
| SLType-TW | 0.039 | 0.003 | 0.054 | 0.086 |
| Category-Left | 0.068 | 0.010 | -0.402 | -0.43 |
| Category-NoLead | 0.051 | 0.020 | -0.2 | -0.233 |
| Category-Right | 0.082 | 0.041 | -0.047 | 0.048 |

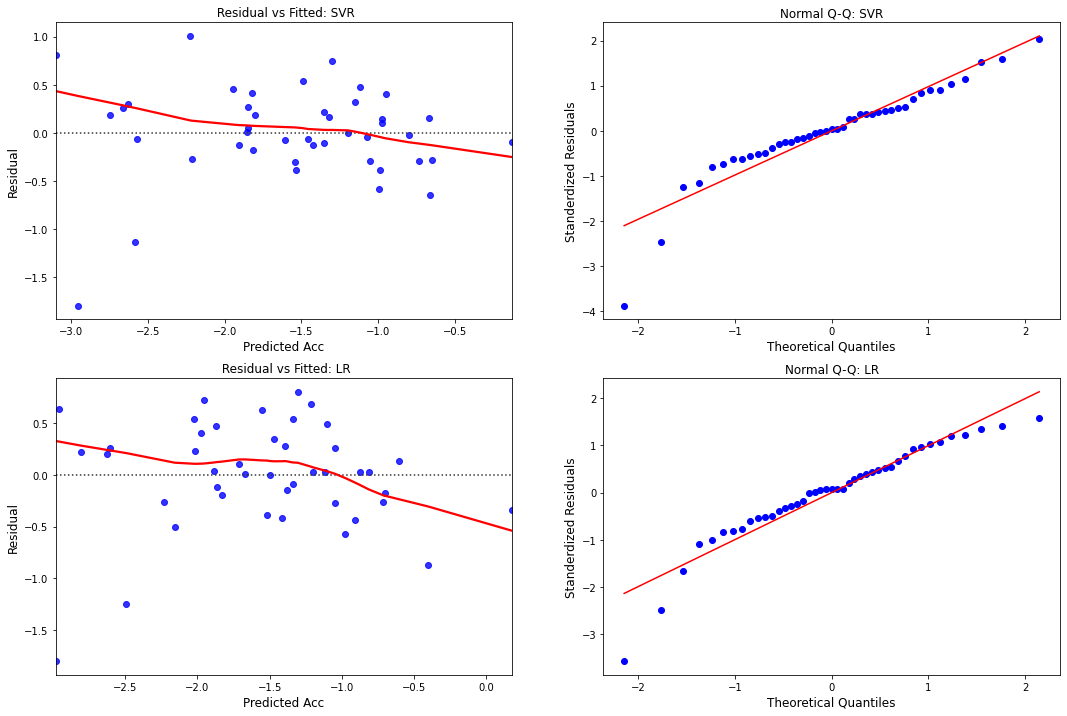
\*\* for SVR and LR option to calculate feature importance is not available, instead coefficients have been displayed.

ANOVA Table:



Residual plots





1. Machine learning models on unbinned data.

* Widening dataset

Model performance on training and test data:

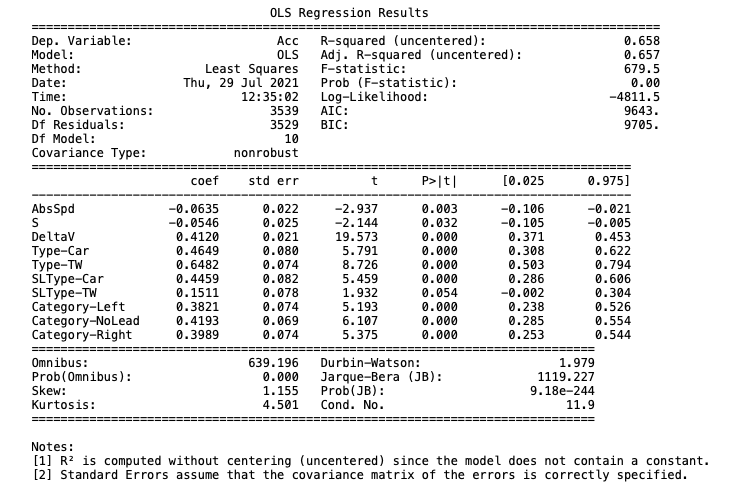
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Train\_  R2Score | Test\_  R2Score | Train\_  RMSE | Test\_  RMSE | Train\_  MAE | Test\_  MAE |
| XGBR | 0.411 | 0.11 | 0.76 | 0.942 | 0.571 | 0.717 |
| RF | 0.251 | 0.155 | 0.857 | 0.918 | 0.658 | 0.712 |
| SVR | 0.045 | 0.066 | 0.968 | 0.965 | 0.704 | 0.7 |
| LR | 0.103 | 0.147 | 0.938 | 0.922 | 0.725 | 0.711 |

Feature Importance:

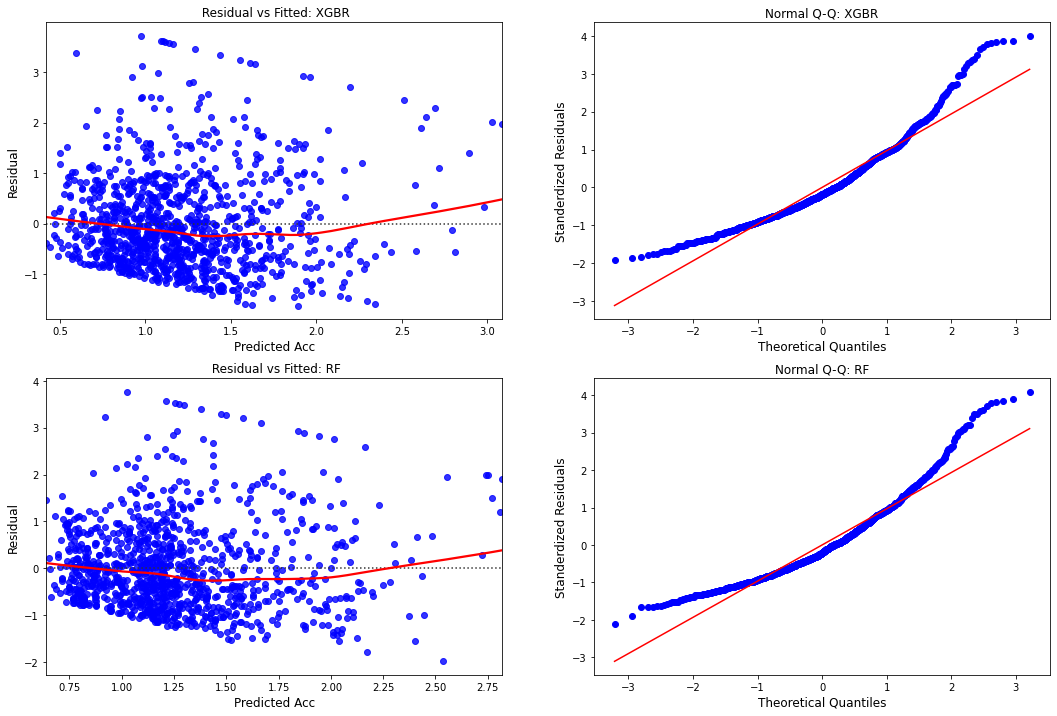
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | XGBR | RF | SVR | LR |
| AbsSpd | 0.079 | 0.182 | 0.031 | -0.022 |
| S | 0.078 | 0.17 | -0.054 | -0.075 |
| DeltaV | 0.126 | 0.563 | 0.311 | 0.398 |
| Type-Car | 0.137 | 0.009 | -0.004 | 0.048 |
| Type-TW | 0.135 | 0.013 | 0.034 | 0.2 |
| SLType-Car | 0.095 | 0.005 | 0.12 | -0.062 |
| SLType-TW | 0.135 | 0.034 | -0.12 | -0.34 |
| Category-Left | 0.072 | 0.004 | -0.016 | 0.076 |
| Category-NoLead | 0.067 | 0.012 | -0.004 | 0.119 |
| Category-Right | 0.076 | 0.009 | 0.007 | 0.121 |

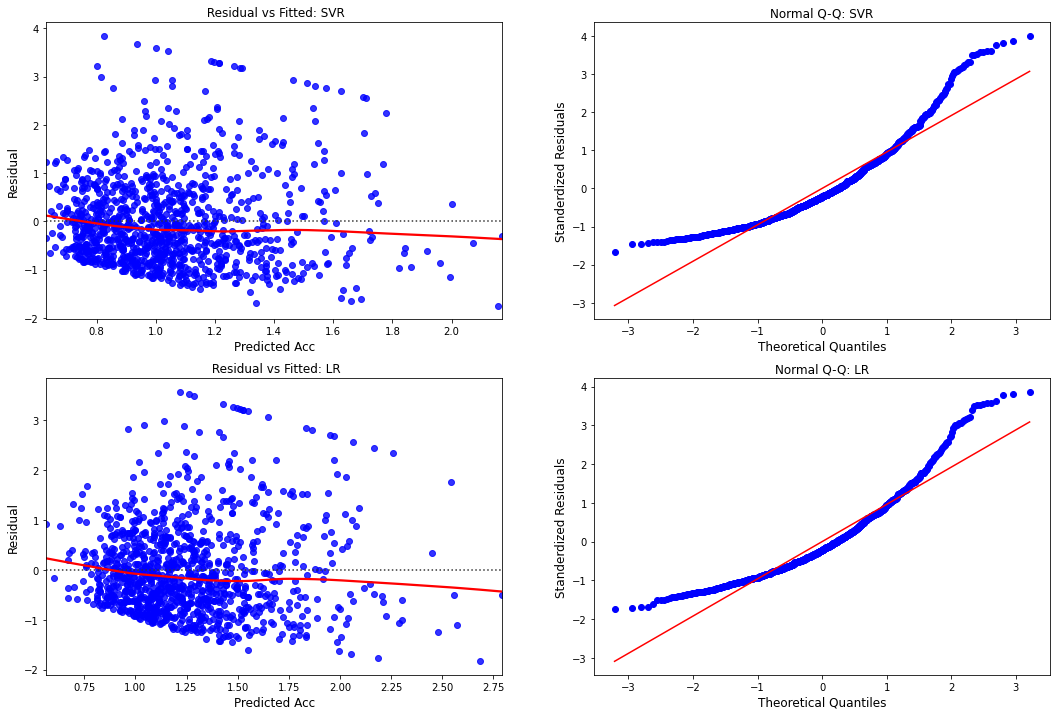
\*\* for SVR and LR option to calculate feature importance is not available, instead coefficients have been displayed.

ANOVA TABLE



Residual plots:





* Narrowing

Model performance on training and test set.

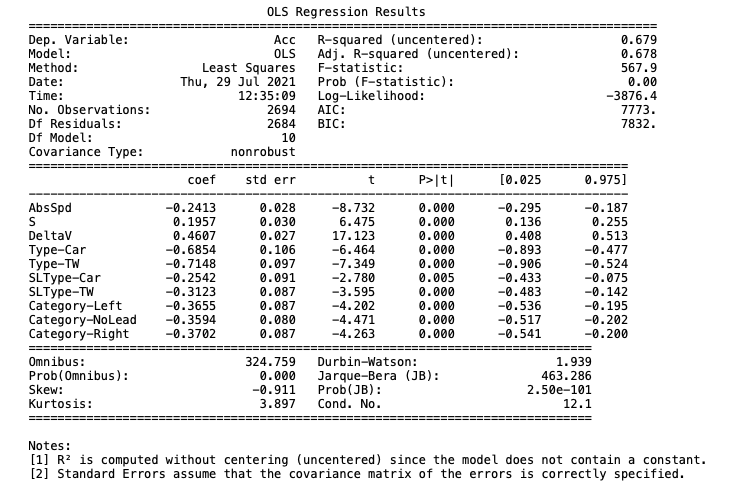
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Train\_  R2Score | Test\_  R2SCore | Train\_  RMSE | Test\_  RMSE | Train\_  MAE | Test\_  MAE |
| XGBR | 0.507 | 0.145 | 0.796 | 0.997 | 0.608 | 0.766 |
| RF | 0.312 | 0.188 | 0.941 | 0.972 | 0.732 | 0.757 |
| SVR | 0.149 | 0.183 | 1.047 | 0.975 | 0.785 | 0.737 |
| LR | 0.177 | 0.196 | 1.029 | 0.967 | 0.8 | 0.757 |

Feature Importance:

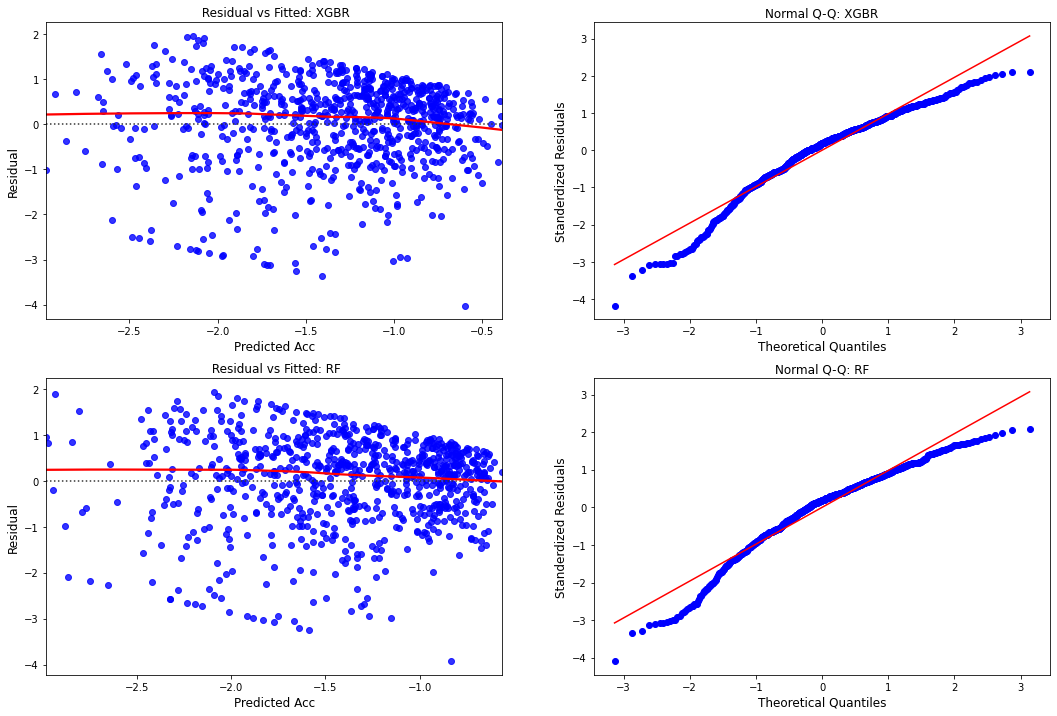
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | XGBR | RF | SVR | LR |
| AbsSpd | 0.107 | 0.198 | -0.225 | -0.226 |
| S | 0.096 | 0.169 | 0.172 | 0.21 |
| DeltaV | 0.197 | 0.603 | 0.487 | 0.489 |
| Type-Car | 0.084 | 0.001 | -0.025 | -0.037 |
| Type-TW | 0.087 | 0.002 | 0.025 | -0.063 |
| SLType-Car | 0.088 | 0.004 | 0.084 | 0.114 |
| SLType-TW | 0.089 | 0.005 | -0.033 | 0.023 |
| Category-Left | 0.089 | 0.002 | -0.094 | -0.126 |
| Category-NoLead | 0.073 | 0.012 | -0.134 | -0.135 |
| Category-Right | 0.089 | 0.004 | -0.148 | -0.126 |

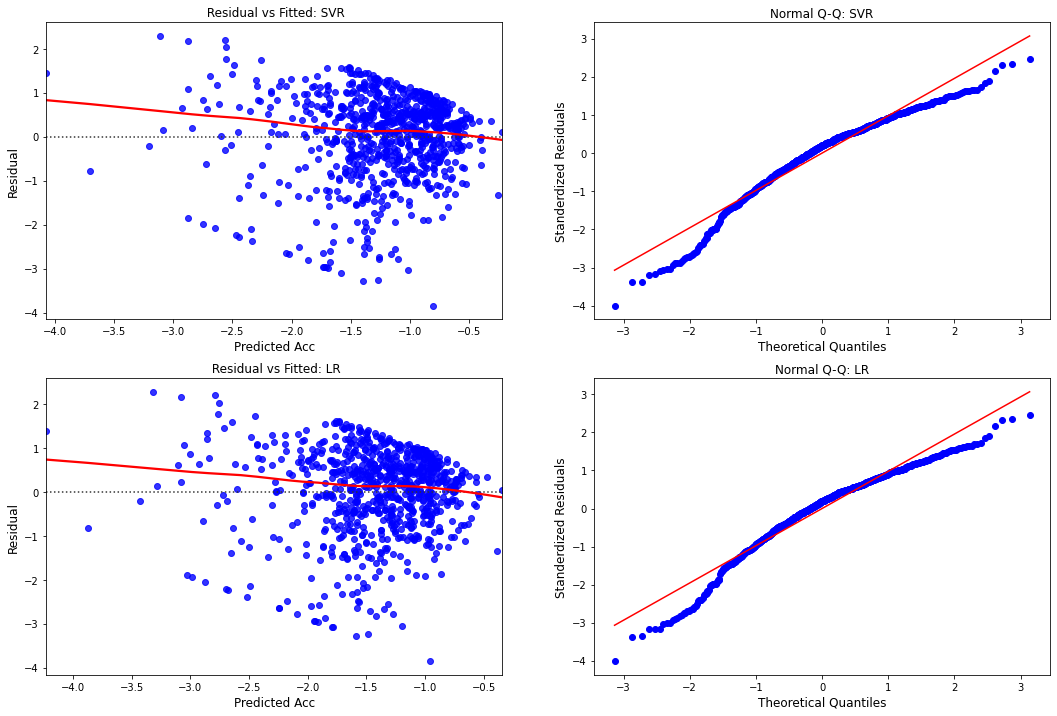
\*\* for SVR and LR option to calculate feature importance is not available, instead coefficients have been displayed.

ANOVA Table:



Residual Plots:





***Comparison of binned and unbinned model on Binned dataset:***

Widening dataset:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Binned dataset | Binned Models | | | Unbinned Models | | |
|  | **R2\_score** | **RMSE** | **MAE** | **R2\_score** | **RMSE** | **MAE** |
| XGBR | 0.476 | 0.362 | 0.262 | -0.336 | 0.578 | 0.416 |
| RF | 0.329 | 0.409 | 0.286 | -0.08 | 0.519 | 0.333 |
| SVR | -0.014 | 0.503 | 0.335 | -0.179 | 0.543 | 0.406 |
| LR | 0.209 | 0.444 | 0.312 | 0.291 | 0.421 | 0.265 |

Narrowing dataset:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Binned dataset | Binned Models | | | Unbinned Models | | |
|  | **R2\_score** | **RMSE** | **MAE** | **R2\_score** | **RMSE** | **MAE** |
| XGBR | 0.647 | 0.516 | 0.353 | 0.277 | 0.738 | 0.477 |
| RF | 0.648 | 0.515 | 0.325 | 0.416 | 0.664 | 0.398 |
| SVR | 0.692 | 0.482 | 0.333 | 0.569 | 0.57 | 0.41 |
| LR | 0.654 | 0.511 | 0.376 | 0.594 | 0.553 | 0.387 |

***Comparison of binned and unbinned model on Unbinned dataset:***

Widening dataset

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Unbinned dataset | Binned Models | | | Unbinned Models | | |
|  | **R2\_score** | **RMSE** | **MAE** | **R2\_score** | **RMSE** | **MAE** |
| XGBR | 0.037 | 0.98 | 0.765 | 0.11 | 0.942 | 0.717 |
| RF | 0.04 | 0.978 | 0.761 | 0.155 | 0.918 | 0.712 |
| SVR | 0.061 | 0.967 | 0.741 | 0.066 | 0.965 | 0.7 |
| LR | 0.058 | 0.969 | 0.733 | 0.147 | 0.922 | 0.711 |

Narrowing dataset

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Unbinned dataset | Binned Models | | | Unbinned Models | | |
|  | **R2\_score** | **RMSE** | **MAE** | **R2\_score** | **RMSE** | **MAE** |
| XGBR | 0.145 | 0.997 | 0.785 | 0.145 | 0.997 | 0.766 |
| RF | 0.152 | 0.993 | 0.786 | 0.188 | 0.972 | 0.757 |
| SVR | 0.161 | 0.988 | 0.787 | 0.183 | 0.975 | 0.737 |
| LR | 0.137 | 1.002 | 0.795 | 0.196 | 0.967 | 0.757 |

***Combined data [Acc Gap widening + Acc Gap Narrowing]***

Model performance on training and test data:

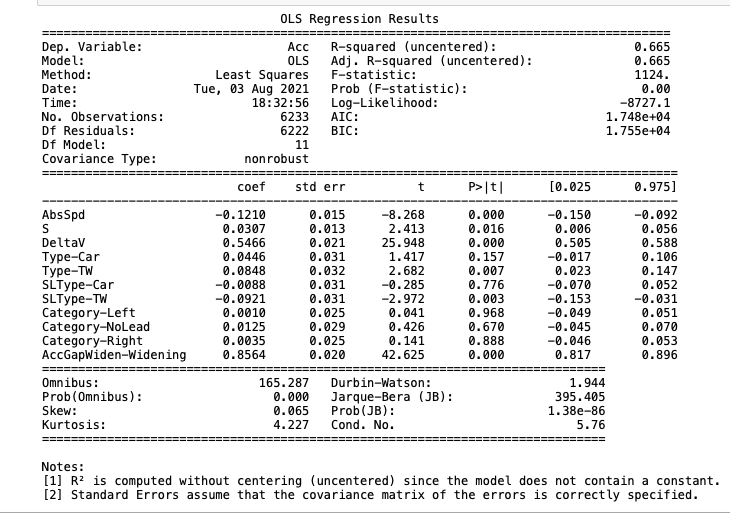
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Train\_  R2Score | Test\_  R2Score | Train\_  RMSE | Test\_  RMSE | Train\_  MAE | Test\_  MAE |
| XGBR | 0.794 | 0.656 | 0.765 | 1.004 | 0.579 | 0.758 |
| RF | 0.723 | 0.68 | 0.886 | 0.969 | 0.684 | 0.749 |
| SVR | 0.646 | 0.655 | 1.001 | 1.007 | 0.739 | 0.739 |
| LR | 0.664 | 0.676 | 0.975 | 0.976 | 0.758 | 0.755 |

Feature Importance:

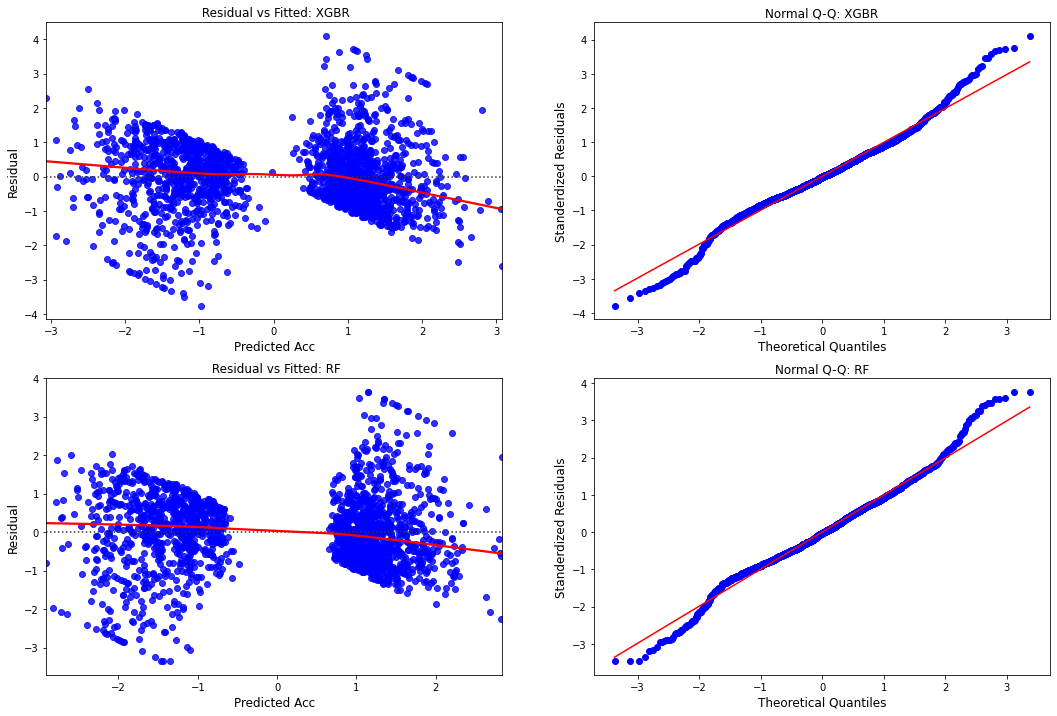
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | XGBR | RF | SVR | LR |
| AbsSpd | 0.004 | 0.032 | -0.095 | -0.124 |
| S | 0.003 | 0.027 | 0.02 | 0.025 |
| DeltaV | 0.024 | 0.637 | 0.433 | 0.526 |
| Type-Car | 0.003 | 0.001 | 0.038 | 0.059 |
| Type-TW | 0.003 | 0.001 | 0.065 | 0.091 |
| SLType-Car | 0.003 | 0.001 | 0.029 | 0.014 |
| SLType-TW | 0.003 | 0.002 | -0.054 | -0.074 |
| Category-Left | 0.003 | 0.001 | -0.024 | 0.016 |
| Category-NoLead | 0.003 | 0.002 | -0.048 | 0.02 |
| Category-Right | 0.003 | 0.001 | -0.015 | 0.018 |
| AccGapWidening | 0.947 | 0.295 | 0.74 | 0.869 |

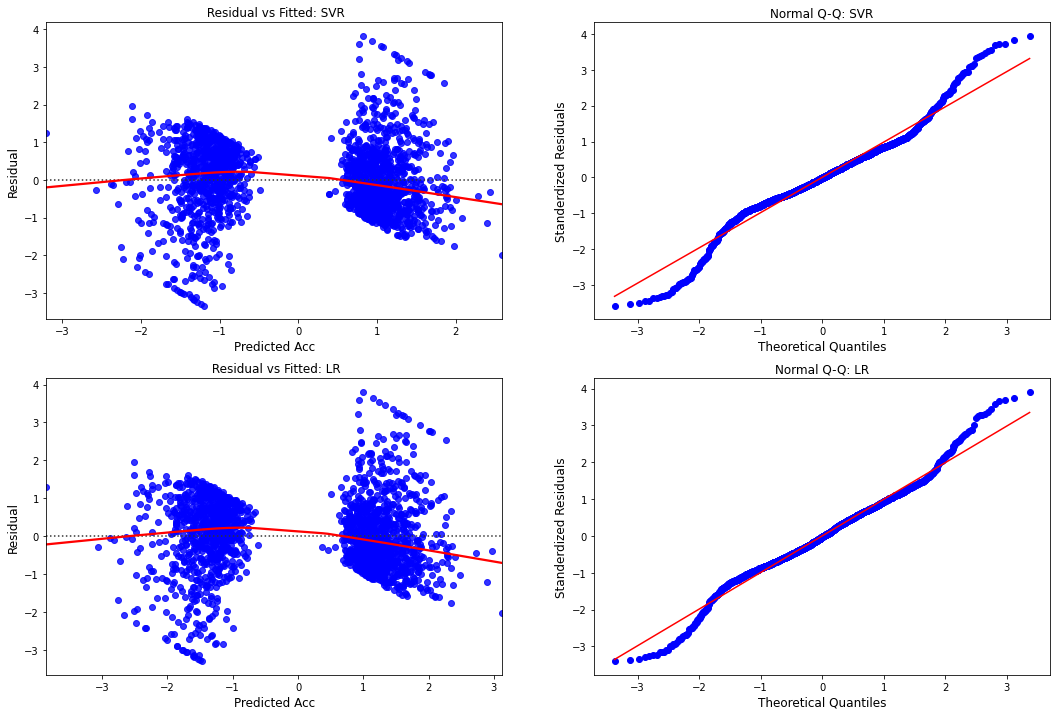
\*\* for SVR and LR option to calculate feature importance is not available, instead coefficients have been displayed.

ANOVA Table:



Residual Plots:





***Combined data [Acc Gap widening + Acc Gap Narrowing] [ Binned data]***

Model performance on training and test data:

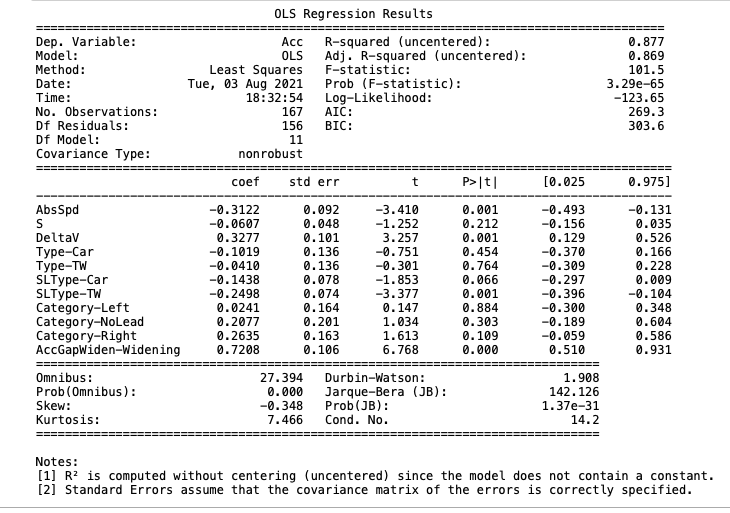
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Train\_  R2Score | Test\_  R2Score | Train\_  RMSE | Test\_  RMSE | Train\_  MAE | Test\_  MAE |
| XGBR | 0.909 | 0.869 | 0.434 | 0.529 | 0.236 | 0.362 |
| RF | 0.869 | 0.876 | 0.52 | 0.514 | 0.307 | 0.347 |
| SVR | 0.863 | 0.891 | 0.533 | 0.482 | 0.324 | 0.317 |
| LR | 0.874 | 0.886 | 0.511 | 0.492 | 0.344 | 0.335 |

Feature Importance:

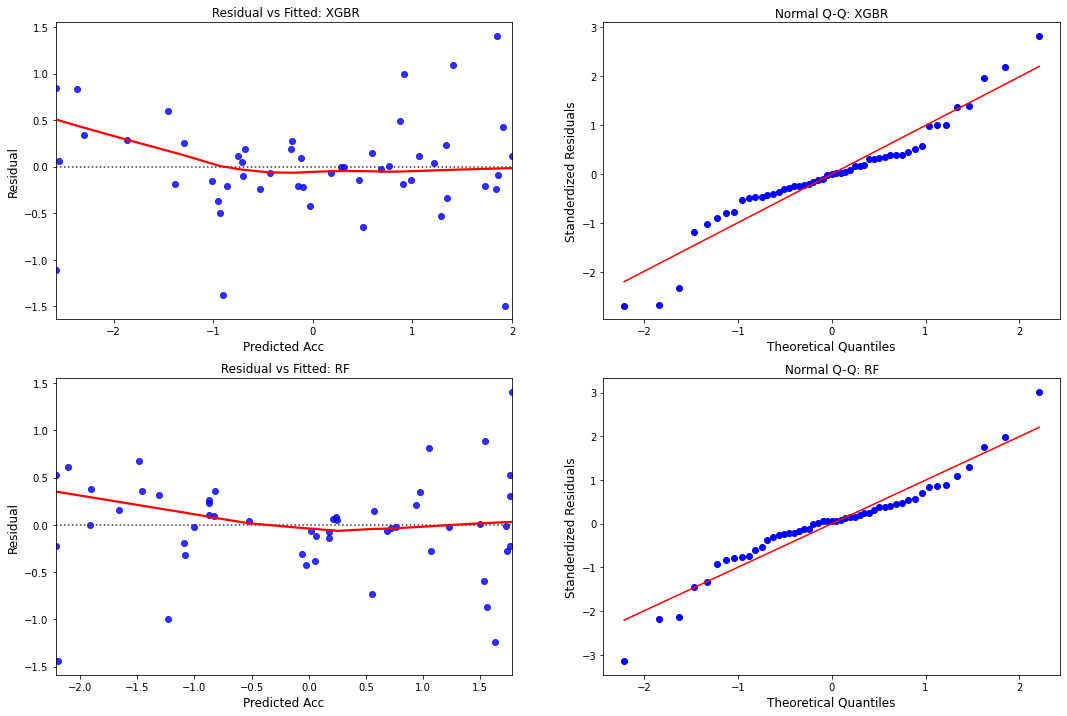
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | XGBR | RF | SVR | LR |
| AbsSpd | 0.081 | 0.054 | -0.16 | -0.288 |
| S | 0.02 | 0.007 | -0.04 | -0.114 |
| DeltaV | 0.525 | 0.603 | 0.497 | 0.349 |
| Type-Car | 0.03 | 0.01 | -0.121 | -0.017 |
| Type-TW | 0.02 | 0.008 | -0.072 | -0.013 |
| SLType-Car | 0.006 | 0.002 | -0.145 | -0.171 |
| SLType-TW | 0.008 | 0 | -0.181 | -0.219 |
| Category-Left | 0.011 | 0.003 | -0.031 | 0.143 |
| Category-NoLead | 0.013 | 0.004 | 0.149 | 0.375 |
| Category-Right | 0.031 | 0.006 | 0.147 | 0.4 |
| AccGapWidening | 0.255 | 0.303 | 0.647 | 0.702 |

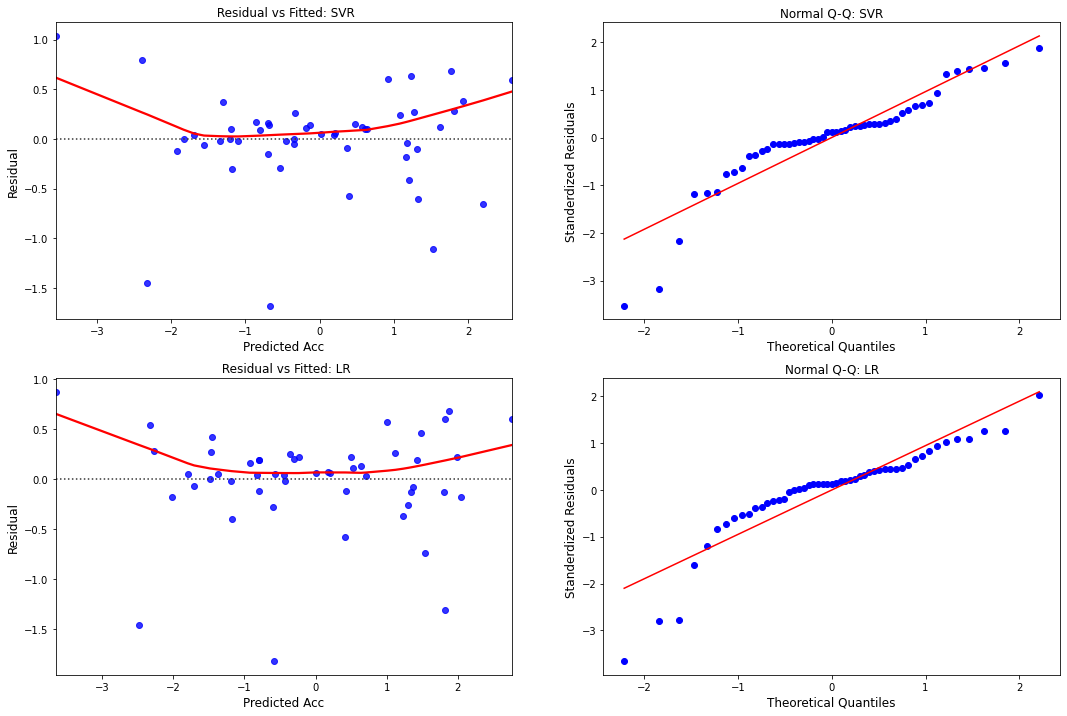
\*\* for SVR and LR option to calculate feature importance is not available, instead coefficients have been displayed.

ANOVA Table:



Residual Plots:





***Comparison of binned and unbinned model on Binned dataset:***

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Binned dataset | Binned Models | | | Unbinned Models | | |
|  | **R2\_score** | **RMSE** | **MAE** | **R2\_score** | **RMSE** | **MAE** |
| XGBR | 0.869 | 0.529 | 0.362 | 0.682 | 0.824 | 0.66 |
| RF | 0.876 | 0.514 | 0.347 | 0.753 | 0.726 | 0.586 |
| SVR | 0.891 | 0.482 | 0.317 | 0.854 | 0.559 | 0.39 |
| LR | 0.886 | 0.492 | 0.335 | 0.856 | 0.554 | 0.386 |

***Comparison of binned and unbinned model on Unbinned dataset:***

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Unbinned dataset | Binned Models | | | Unbinned Models | | |
|  | **R2\_score** | **RMSE** | **MAE** | **R2\_score** | **RMSE** | **MAE** |
| XGBR | 0.61 | 1.07 | 0.777 | 0.656 | 1.004 | 0.758 |
| RF | 0.644 | 1.022 | 0.763 | 0.68 | 0.969 | 0.749 |
| SVR | 0.631 | 1.041 | 0.771 | 0.655 | 1.007 | 0.739 |
| LR | 0.594 | 1.092 | 0.816 | 0.676 | 0.976 | 0.755 |

***Comparison of combined and individual model:***

Below tabulated results are model performances on Test data for combined dataset and split data [ Widening & Narrowing].

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Widening | | | Narrowing | | | Combined | | |
|  | **R2** | **RMSE** | **MAE** | **R2** | **RMSE** | **MAE** | **R2** | **RMSE** | **MAE** |
| XGBR | 0.11 | 0.942 | 0.717 | 0.145 | 0.997 | 0.766 | 0.656 | 1.005 | 0.774 |
| RF | 0.155 | 0.918 | 0.712 | 0.188 | 0.972 | 0.757 | 0.68 | 0.969 | 0.749 |
| SVR | 0.066 | 0.965 | 0.7 | 0.183 | 0.975 | 0.737 | 0.655 | 1.007 | 0.739 |
| LR | 0.147 | 0.922 | 0.711 | 0.196 | 0.967 | 0.757 | 0.676 | 0.976 | 0.755 |

­

Hyperparameters:

***For binned datasets:***

Random Forest:

{'bootstrap': True,

'ccp\_alpha': 0.0,

'criterion': 'mse',

'max\_depth': 15,

'max\_features': 'auto',

'max\_leaf\_nodes': None,

'max\_samples': None,

'min\_impurity\_decrease': 0.0,

'min\_impurity\_split': None,

'min\_samples\_leaf': 10,

'min\_samples\_split': 10,

'min\_weight\_fraction\_leaf': 0.0,

'n\_estimators': 50,

'n\_jobs': None,

'oob\_score': False,

'random\_state': 10,

'verbose': 0,

'warm\_start': False}

XGBoost Regressor:

{'objective': 'reg:squarederror',

'base\_score': 0.5,

'booster': 'gbtree',

'colsample\_bylevel': 1,

'colsample\_bynode': 1,

'colsample\_bytree': 0.8,

'gamma': 0,

'gpu\_id': -1,

'importance\_type': 'gain',

'interaction\_constraints': '',

'learning\_rate': 0.1,

'max\_delta\_step': 0,

'max\_depth': 15,

'min\_child\_weight': 15,

'missing': nan,

'monotone\_constraints': '()',

'n\_estimators': 50,

'n\_jobs': 4,

'num\_parallel\_tree': 1,

'random\_state': 15,

'reg\_alpha': 0,

'reg\_lambda': 1,

'scale\_pos\_weight': 1,

'subsample': 0.8,

'tree\_method': 'exact',

'validate\_parameters': 1,

'verbosity': None}

***For unbinned datasets:***

Random Forest:

{'bootstrap': True,

'ccp\_alpha': 0.0,

'criterion': 'mse',

'max\_depth': 15,

'max\_features': 'auto',

'max\_leaf\_nodes': None,

'max\_samples': None,

'min\_impurity\_decrease': 0.0,

'min\_impurity\_split': None,

'min\_samples\_leaf': 20,

'min\_samples\_split': 20,

'min\_weight\_fraction\_leaf': 0.0,

'n\_estimators': 100,

'n\_jobs': None,

'oob\_score': False,

'random\_state': 10,

'verbose': 0,

'warm\_start': False}

XGBoost Regressor:

{'objective': 'reg:squarederror',

'base\_score': 0.5,

'booster': 'gbtree',

'colsample\_bylevel': 1,

'colsample\_bynode': 1,

'colsample\_bytree': 0.8,

'gamma': 0,

'gpu\_id': -1,

'importance\_type': 'gain',

'interaction\_constraints': '',

'learning\_rate': 0.2,

'max\_delta\_step': 0,

'max\_depth': 15,

'min\_child\_weight': 18,

'missing': nan,

'monotone\_constraints': '()',

'n\_estimators': 50,

'n\_jobs': 4,

'num\_parallel\_tree': 1,

'random\_state': 0,

'reg\_alpha': 0,

'reg\_lambda': 1,

'scale\_pos\_weight': 1,

'subsample': 0.84,

'tree\_method': 'exact',

'validate\_parameters': 1,

'verbosity': None}