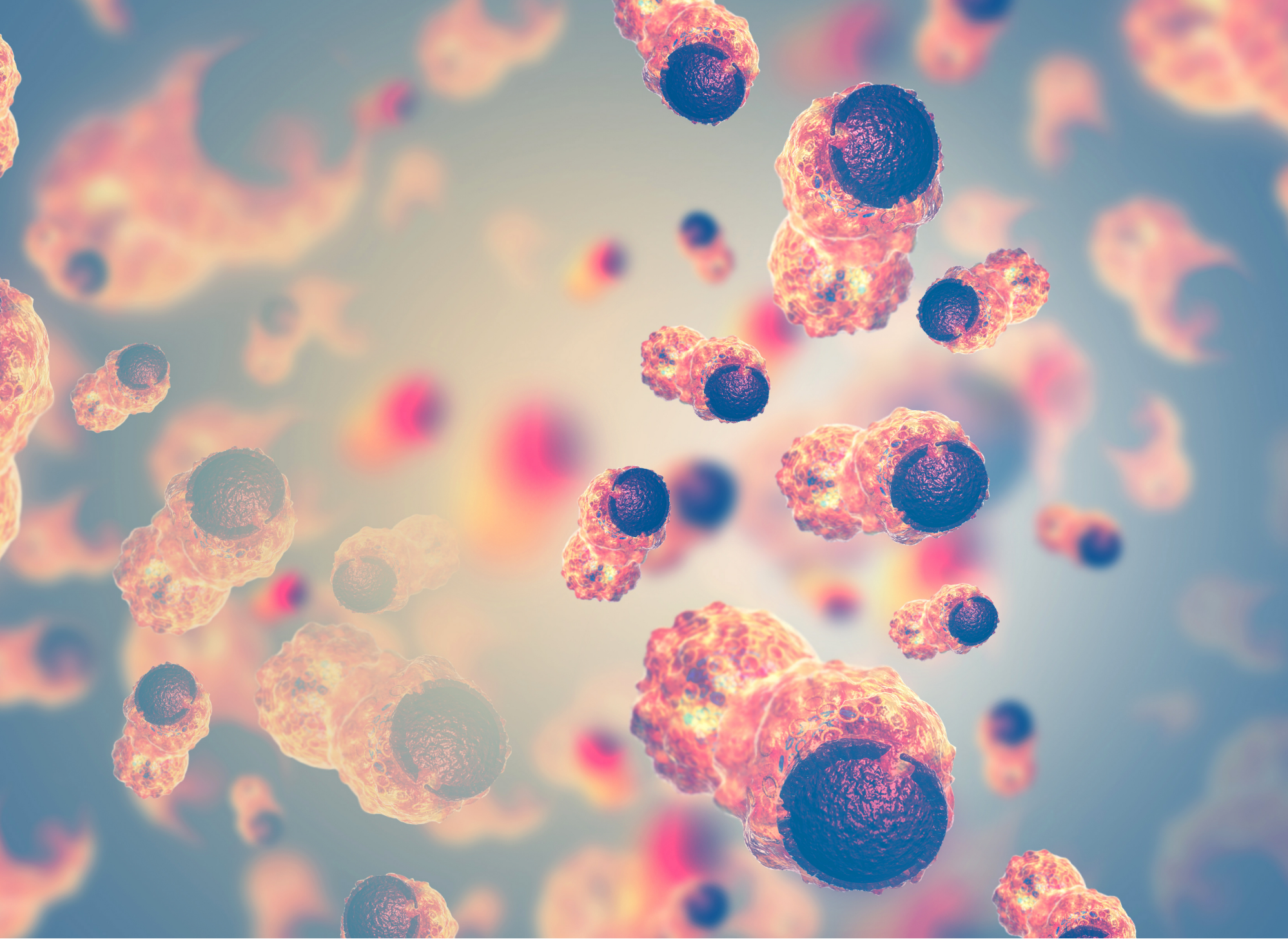


BREAST CANCER SURVIVAL MODELS

Rich Gorham

November 12, 2018

PURPOSE

The purpose of this analysis is the determine the factors that contribute to mortality after a breast cancer diagnosis. Such analysis may be used uncover treatment strategies and provide a baseline for effectiveness.

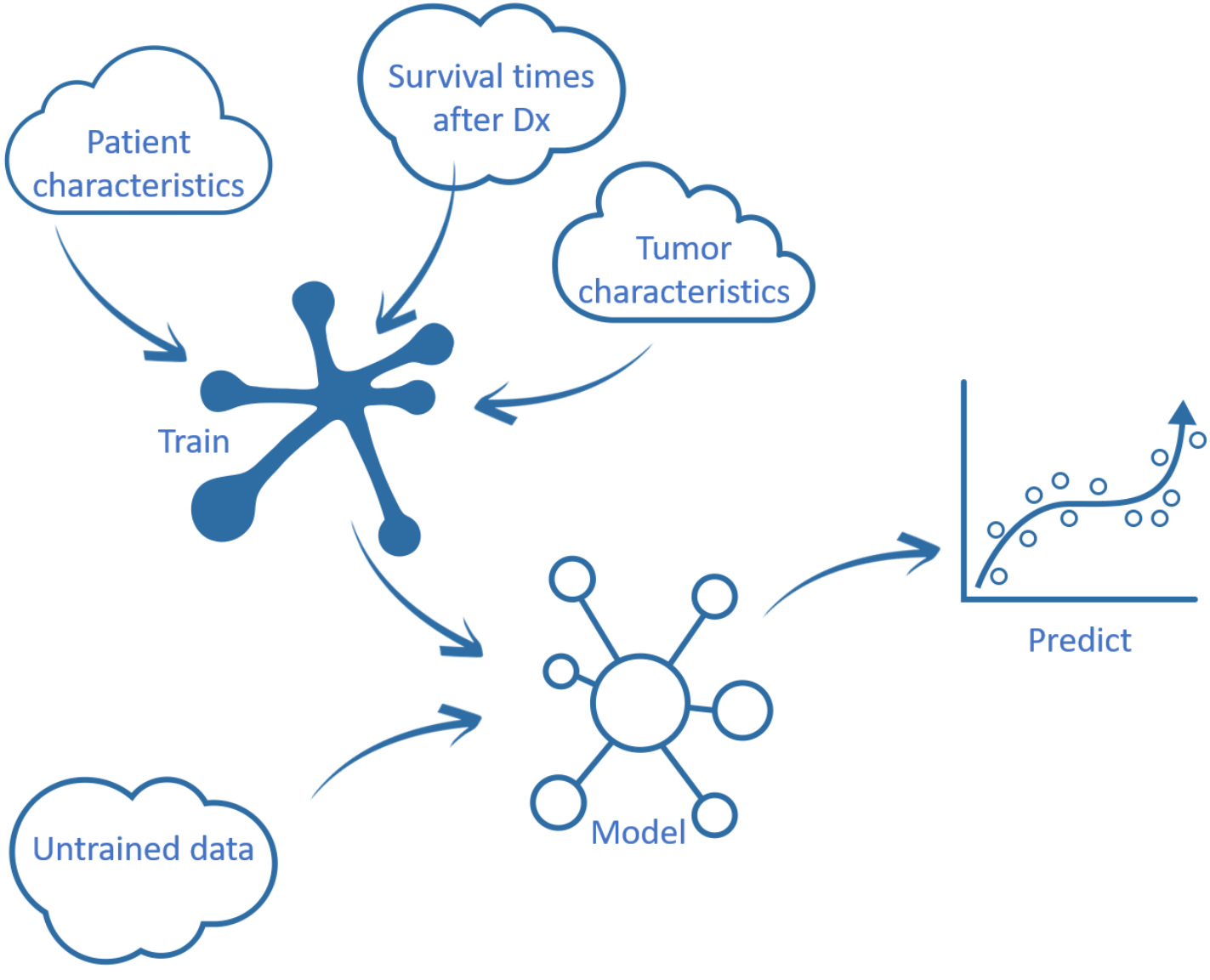
DATA SOURCE

U.S. Department of Health and Human Services, National Institutes of Health, National Cancer Institute

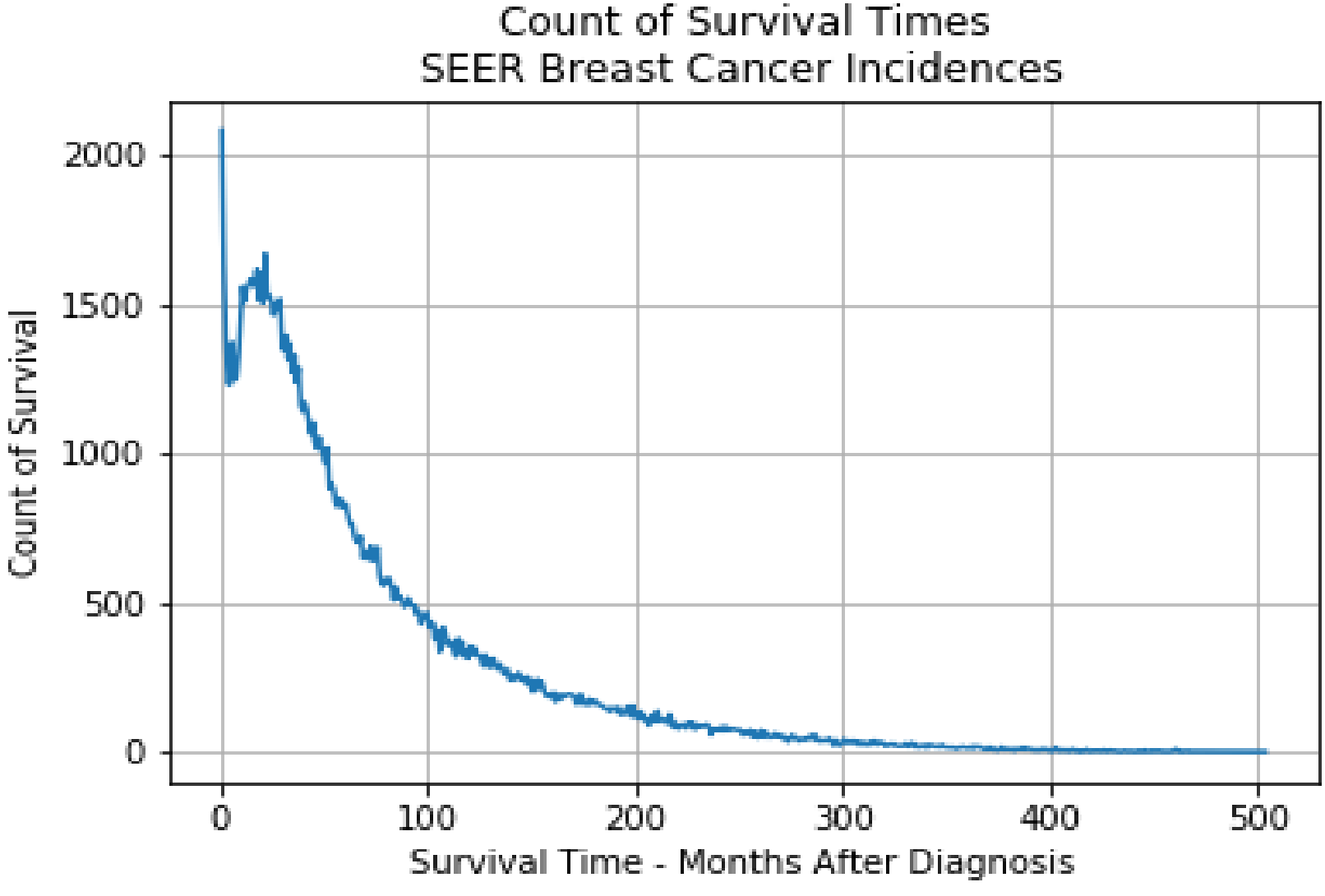
Surveillance, Epidemiology, and End Results (SEER) Program

https://seer.cancer.gov/

APPROACH


Patient characteristics
Survival times after Dx
Tumor characteristics
Train
Untrained data
Model
Predict

DEPENDANT VARIABLE



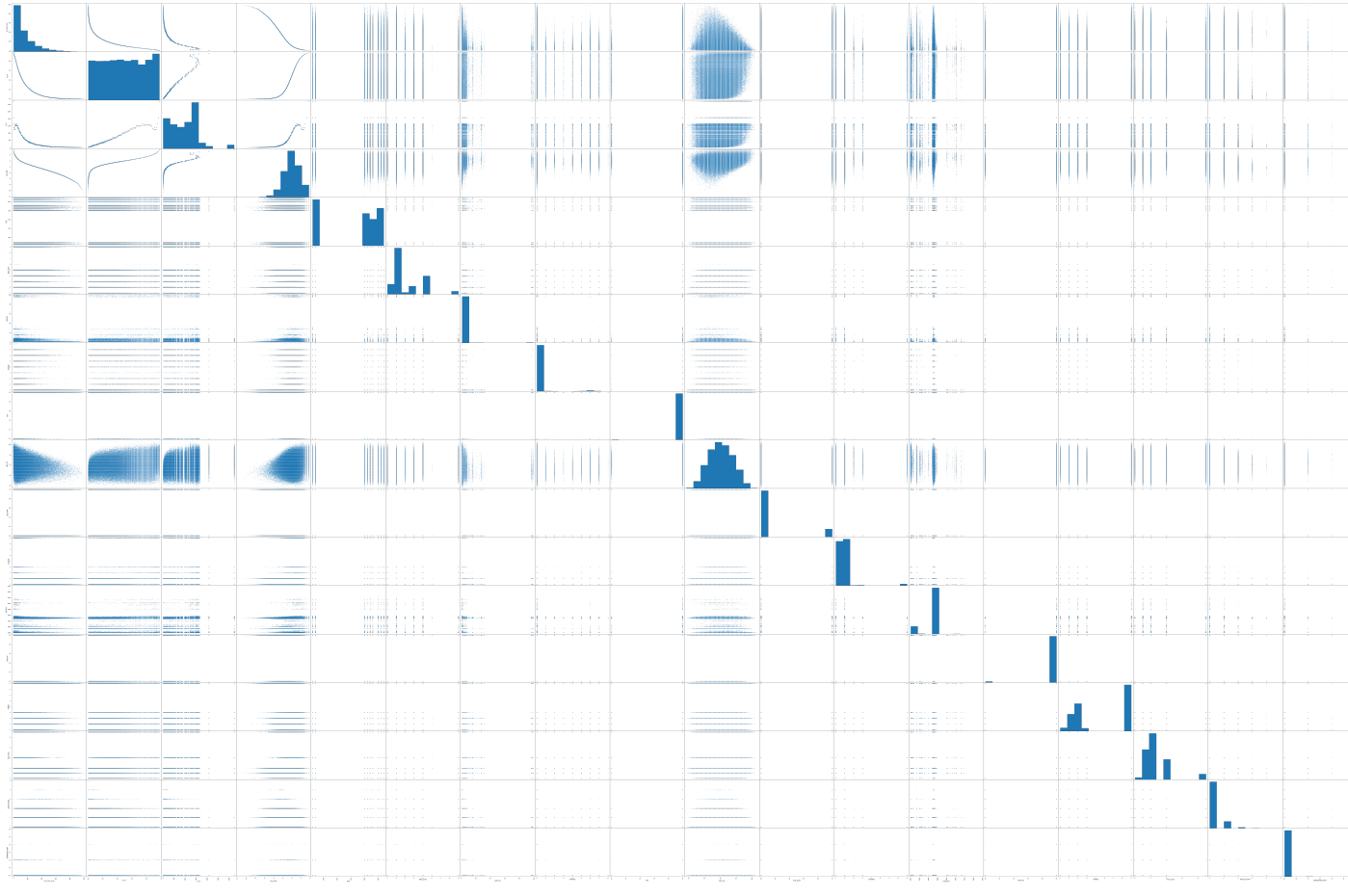
INDEPENDENT VARIABLES - PATIENT CHARACTERISTICS

Variable	Description
REG	Region - US region of diagnosis
MAR_STAT	Marital status of the patient
RACE1V	Ethnicity of the patient
NHIADE	Further classification for hipanic ethnicities
SEX	Gender of the patient
AGE_DX	Age of the patient at diagnosis

INDEPENDENT VARIABLES - NEOPLASM CHARACTERISTICS

Variable	Description
SEQ_NUM	Number of previous diagnoses
LATERAL	Side of body of the diagnosis
HISTO3V	Histologic characteristics of the diagnosis
BEHO3V	Behavior characterization of the neoplasm
GRADE	Classification of the severity of the neoplasm
HST_STGA	Histologic stage of the neoplasm
MALIGCOUNT	Number of malignant neoplasms at diagnosis
BENBORDCOUNT	Number of benign neoplasms at diagnosis
PRIMSITE	Primary site of the diagnosis

MATRIX PLOT

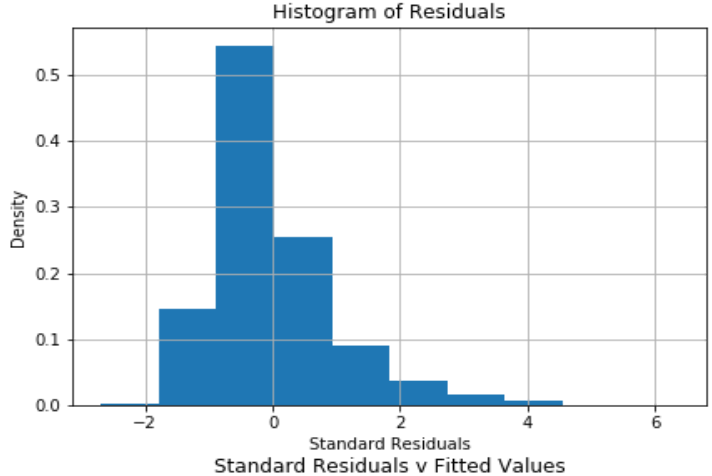


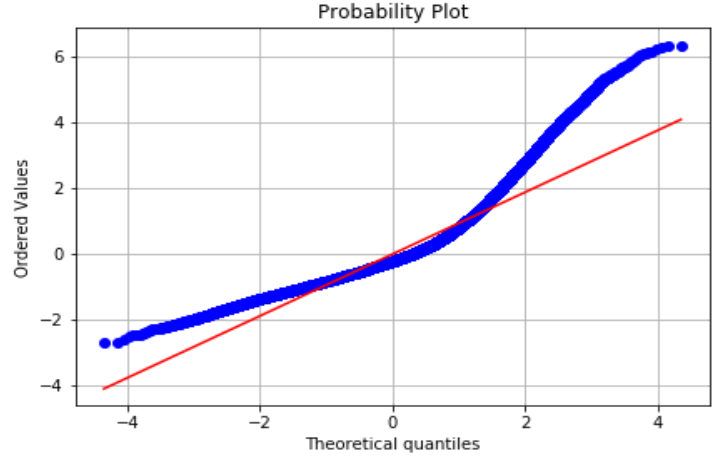
MODEL DIAGNOSITICS ON MONTHS OF SURVIVAL

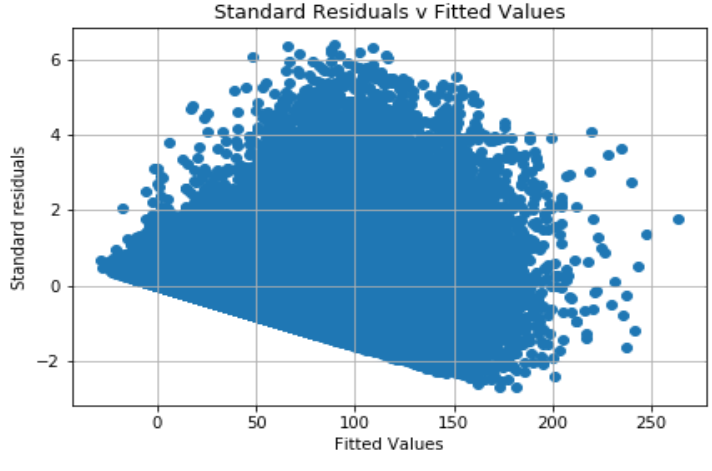
LASSO WITH CROSS VALIDATION AND WEIGHT LEAST SQUARES

Data Set	R^2	Residuals sd	Relative Residual Error
Training	0.24	62.66	0.87
Testing	0.24	62.72	0.87
Weighted Least Squares	0.34	71.96	1.0

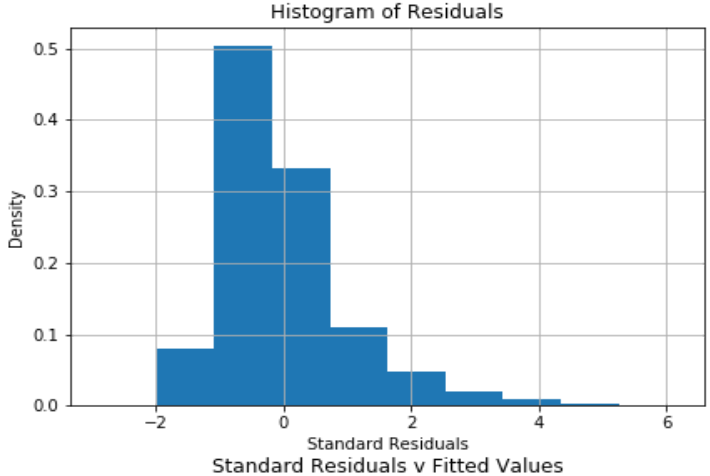
Residual Analysis - Linear
LASSO / Cross Validation
Training Set

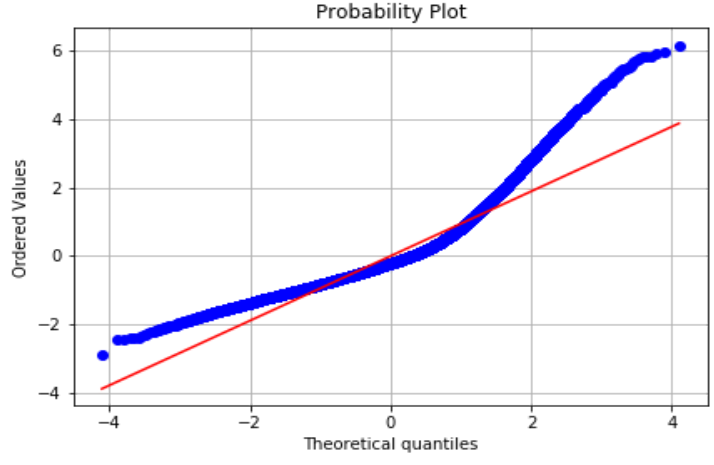


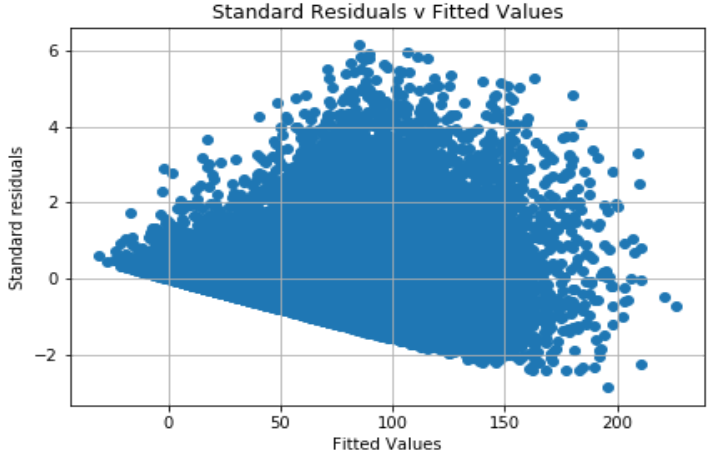




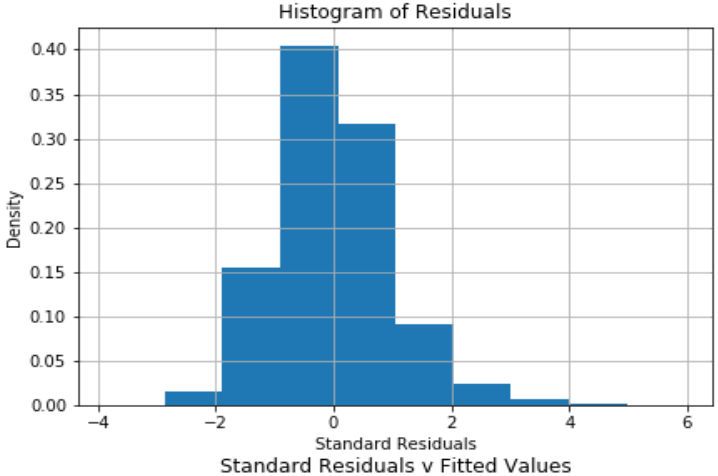
Residual Analysis
LASSO / Cross Validation
Training Set

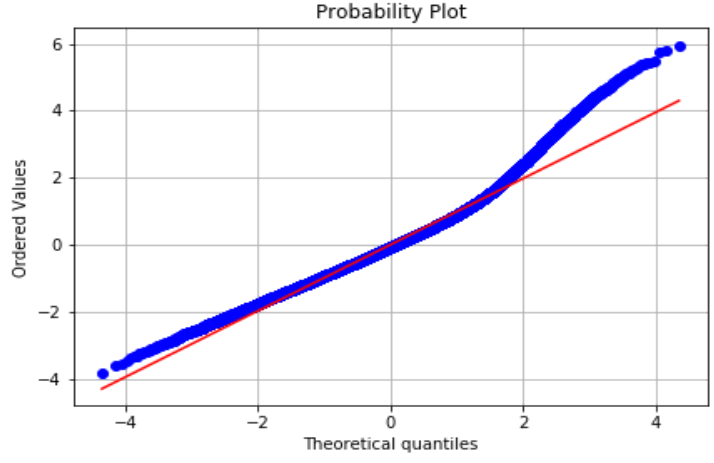


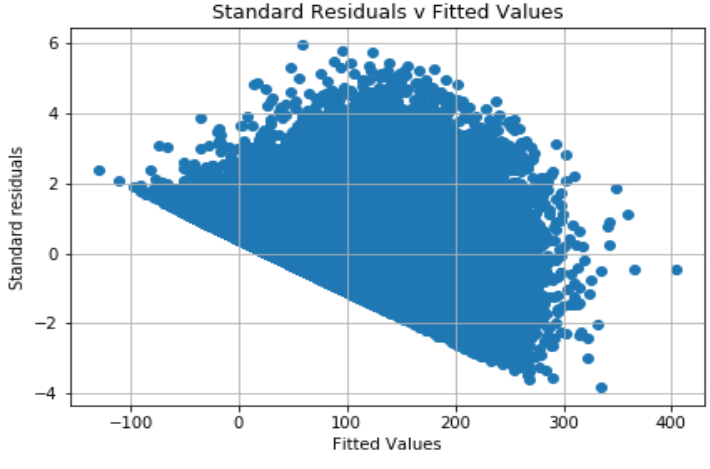




Residual Analysis - Linear
Weighted Least Squares





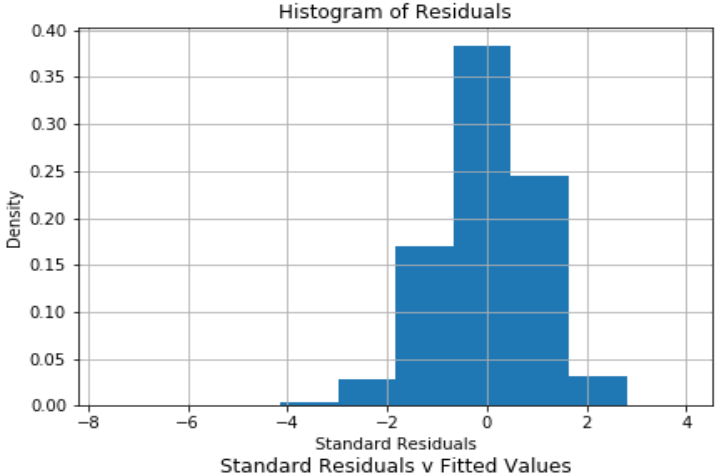


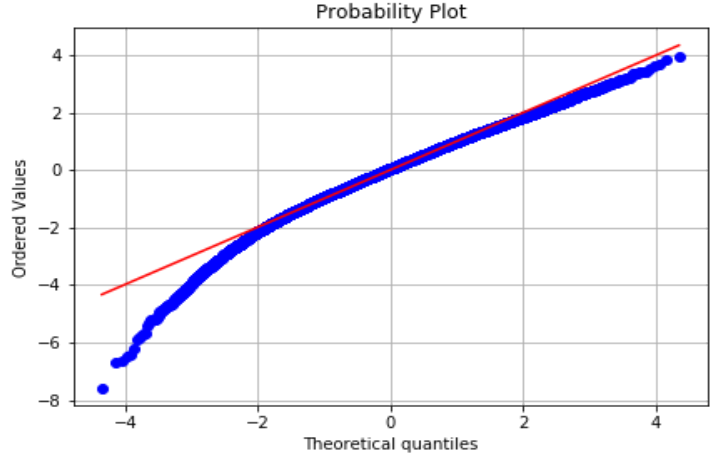
MODEL DIAGNOSITICS ON LOG ODDS

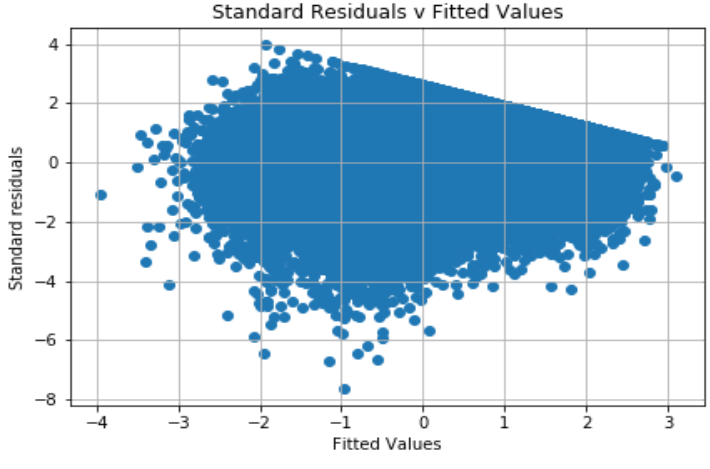
LASSO WITH CROSS VALIDATION AND WEIGHT LEAST SQUARES

Data Set	R^2	Residuals sd	Relative Residual Error
Training	0.32	1.43	0.82
Testing	0.32	1.43	0.83
Weighted Least Squares	0.24	1.87	1.08

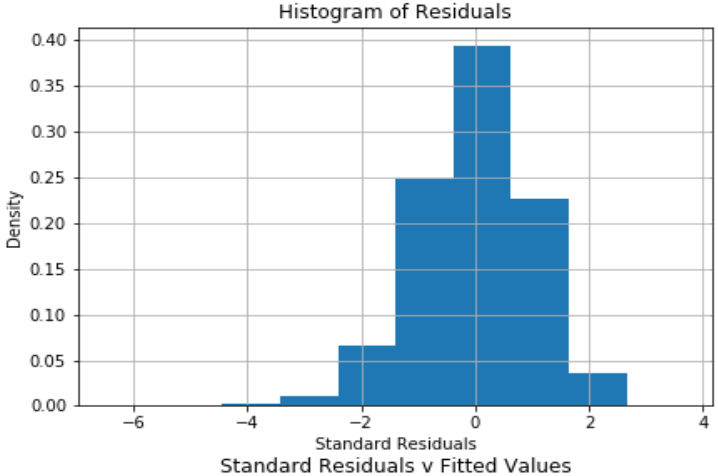
Residual Analysis - Linear
LASSO Cross Validation

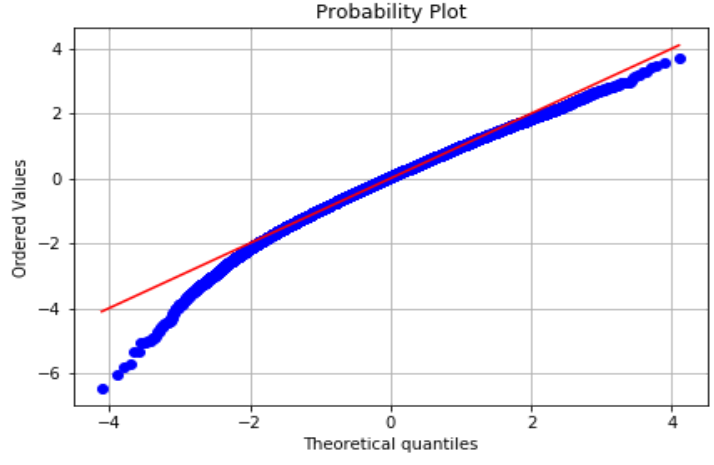


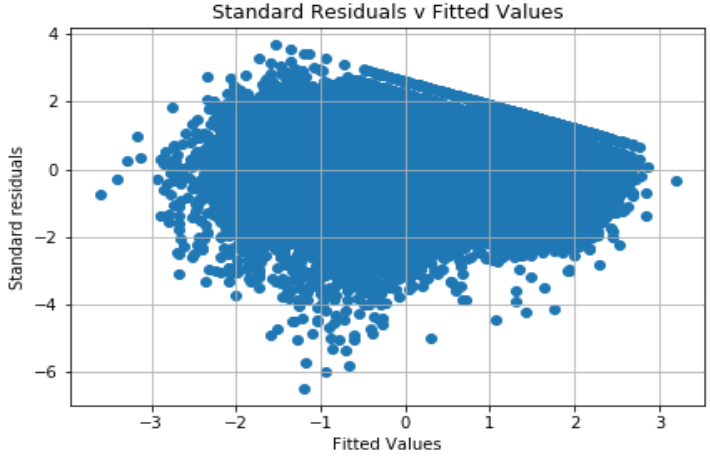




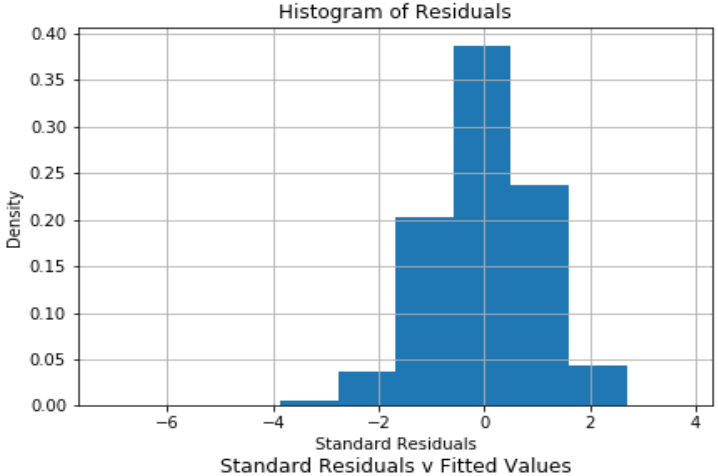
Residual Analysis
LASSO

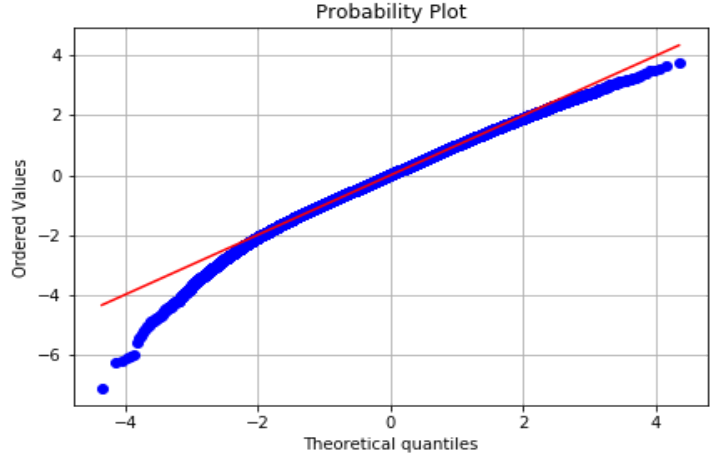


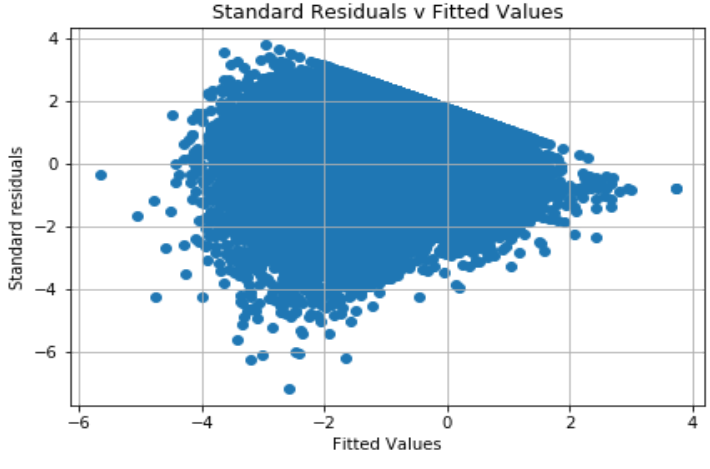




Residual Analysis - Linear
Weighted Least Squares







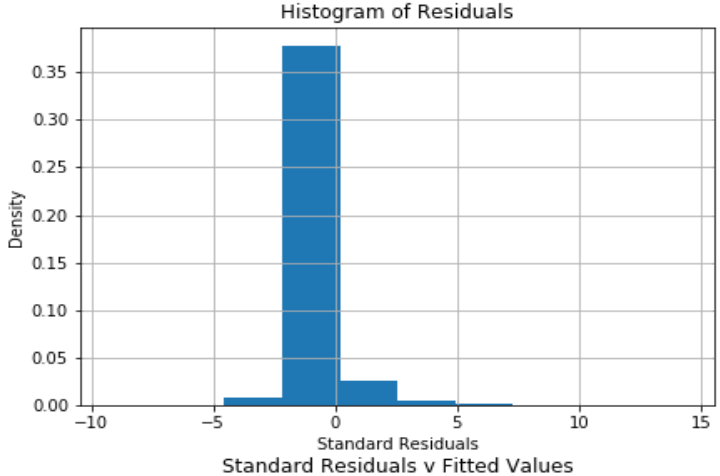
MODEL DIAGNOSITICS ON MONTHS OF SURVIVAL

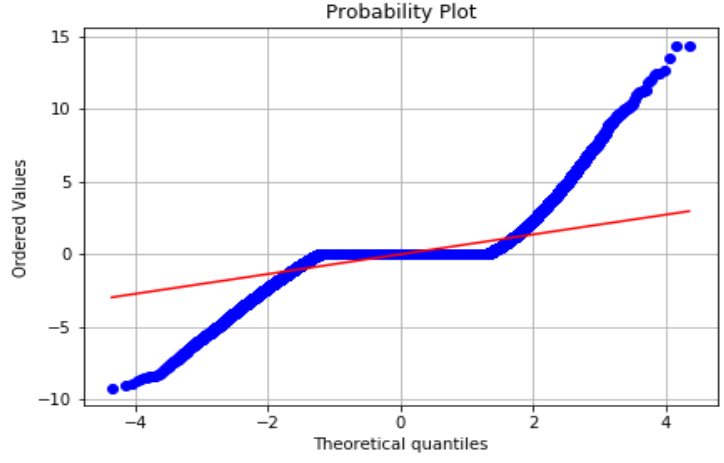
DECISION TREES

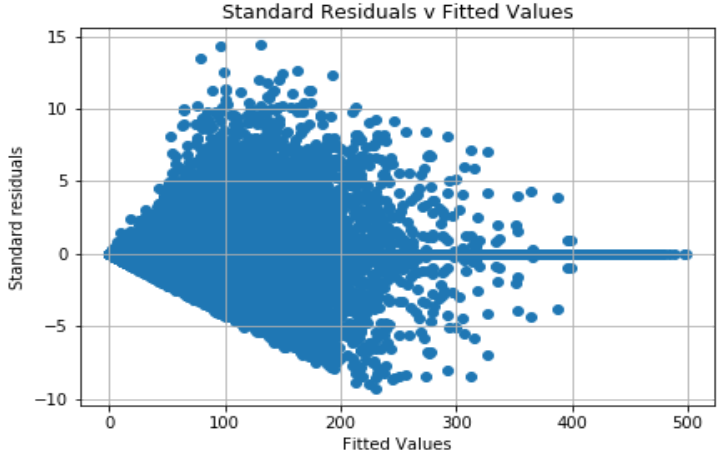
Data Set	R^2	Residuals sd	Relative Residual Error
Training	0.90	23.0	0.32
Testing	-0.39	85.03	1.18

No viable tree identified

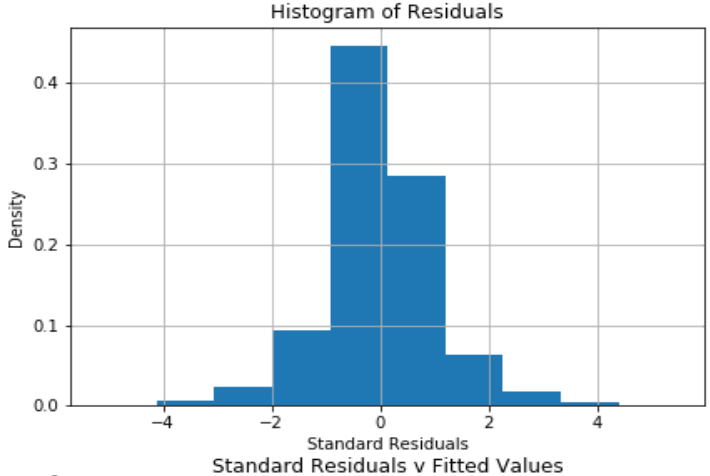
Residual Analysis - Linear
Decision Tree
Training Set

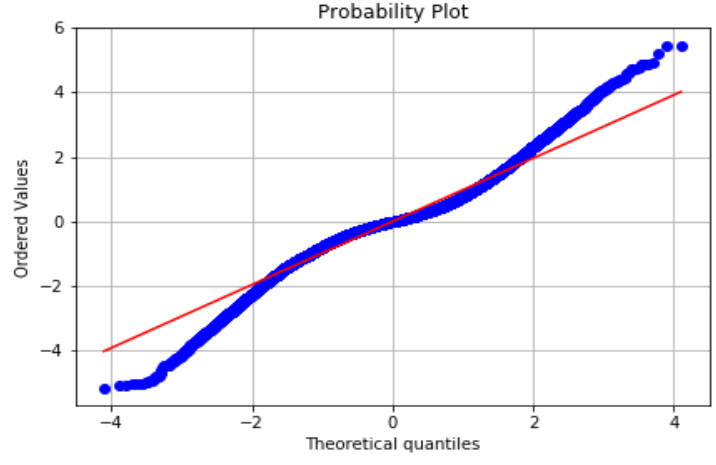


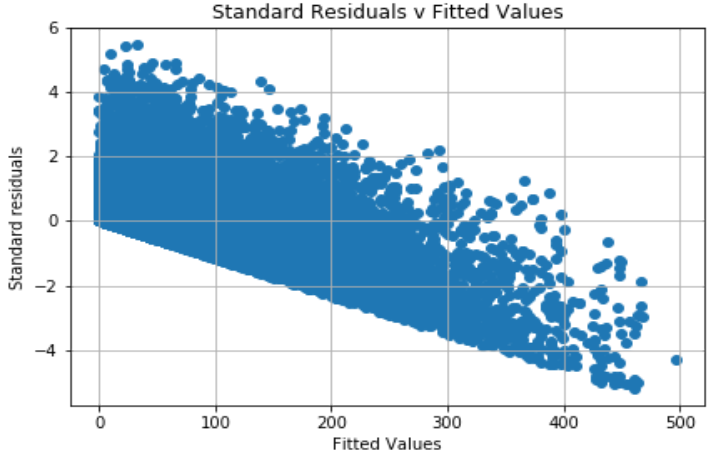




Residual Analysis - Linear
Decision Tree
Training Set







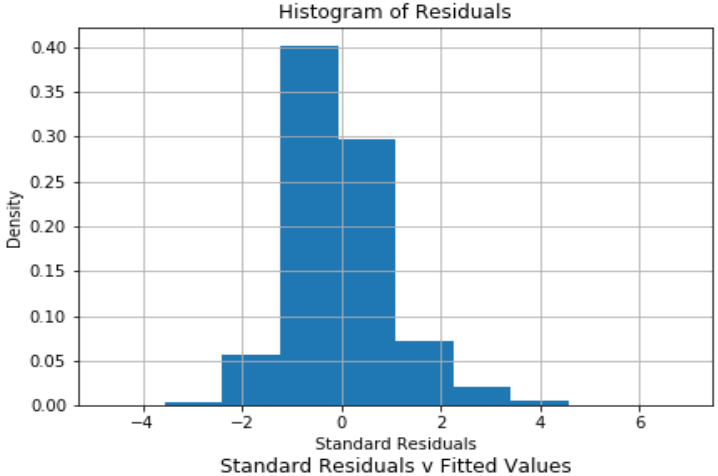
MODEL DIAGNOSITICS ON MONTHS OF SURVIVAL

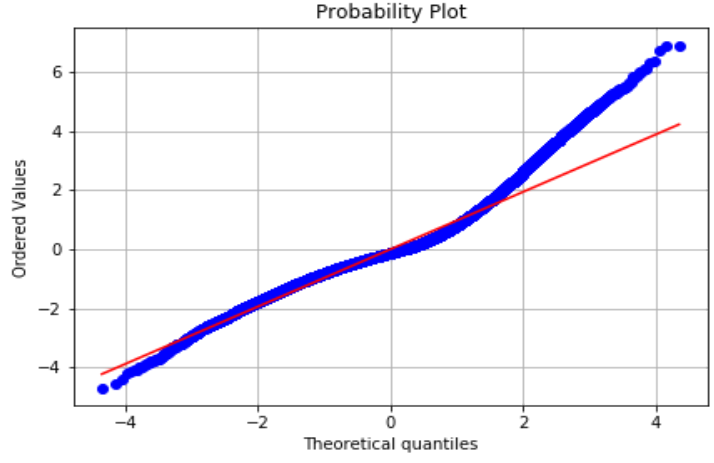
K NEAREST NEIGHBORS

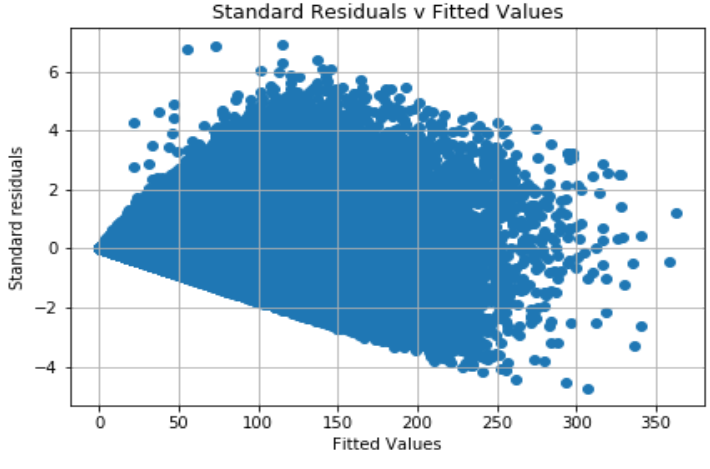
Data Set	R^2	Residuals sd	Relative Residual Error
Training	0.41	55.1	0.77
Testing	0.13	67.84	0.93

No solution to overfitting identified

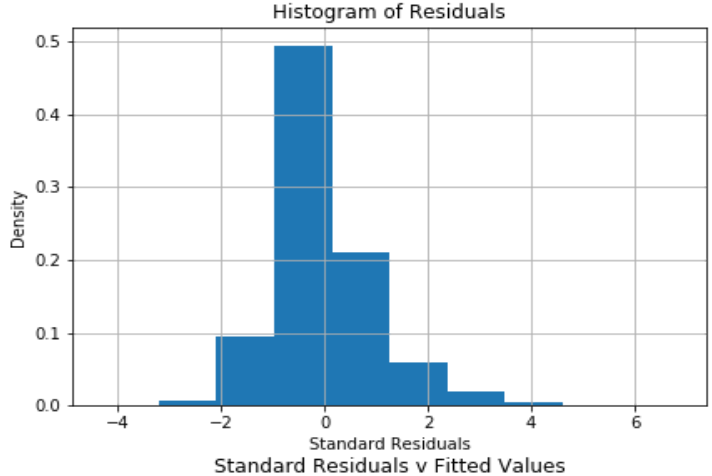
Residual Analysis
K-Nearst neighbors
Training set

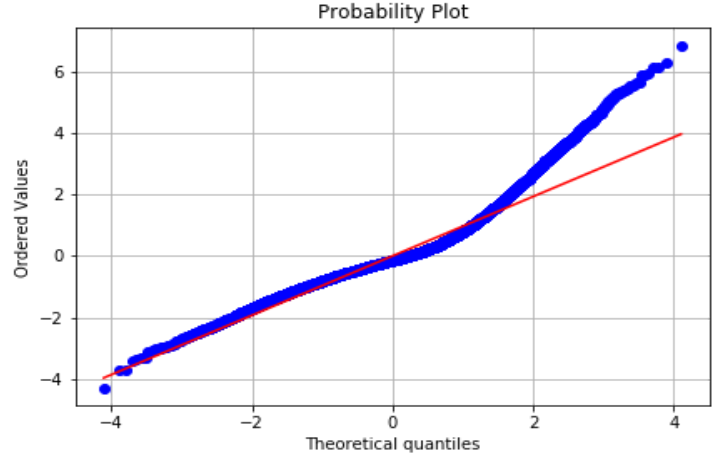


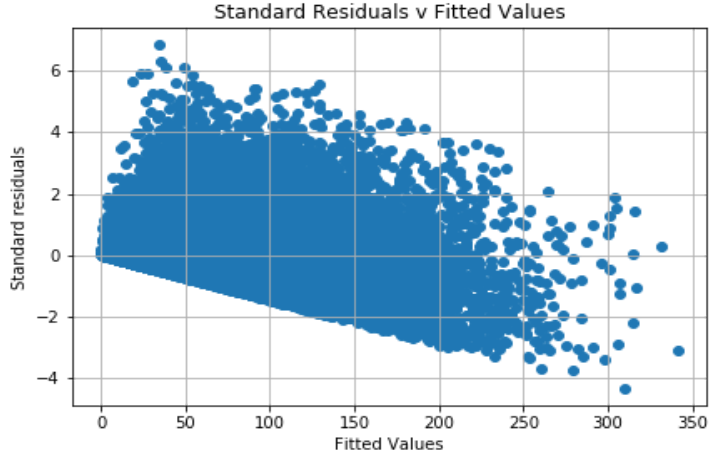




Residual Analysis
k-Nearest neighborsTraining Set



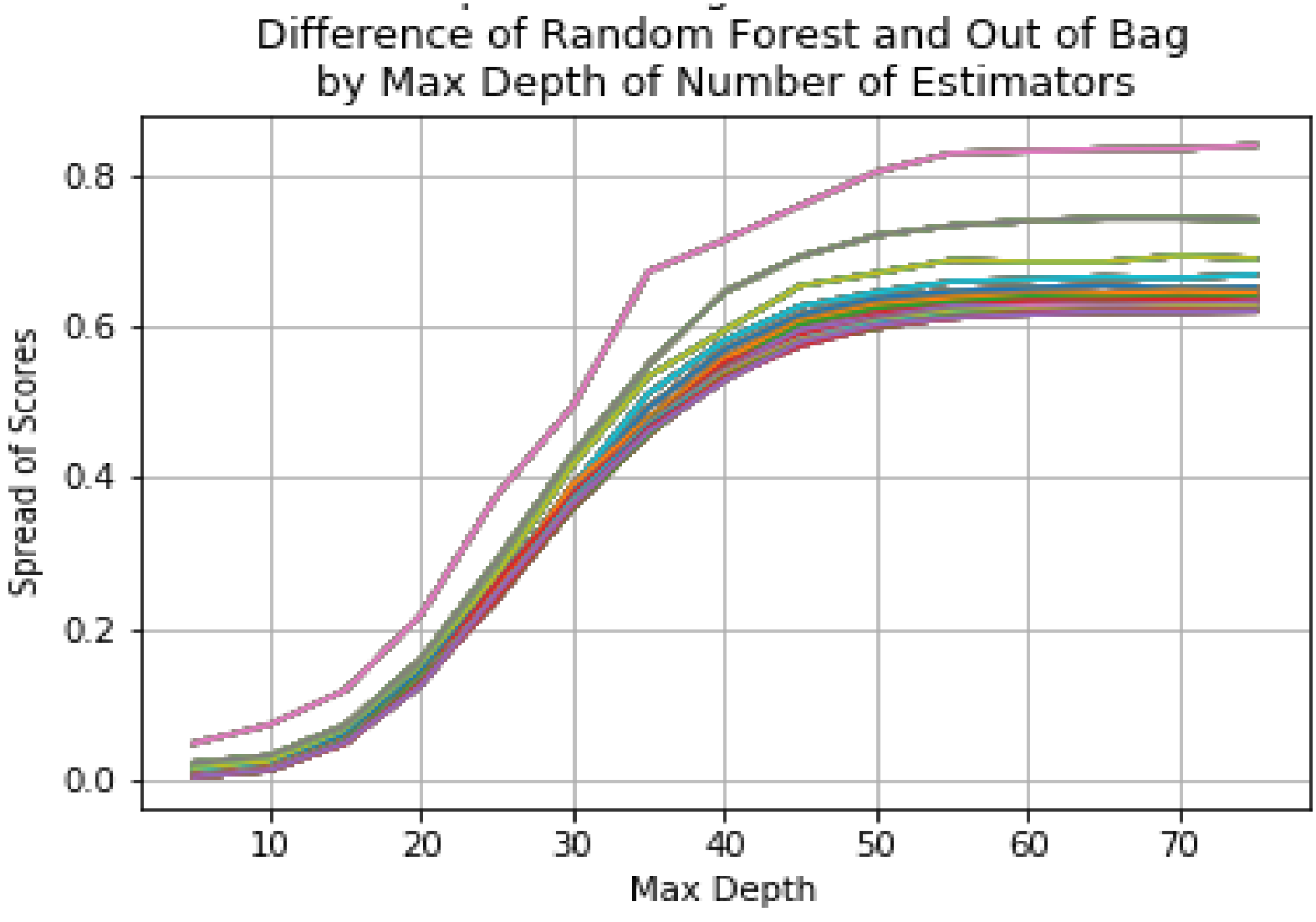




MODEL DIAGNOSITICS ON MONTHS OF SURVIVAL

RANDOM FOREST

Data Set	R^2	Residuals sd	Relative Residual Error
Training	0.83	0.70	0.41
Out of bag	0.22	considerable overfitting	

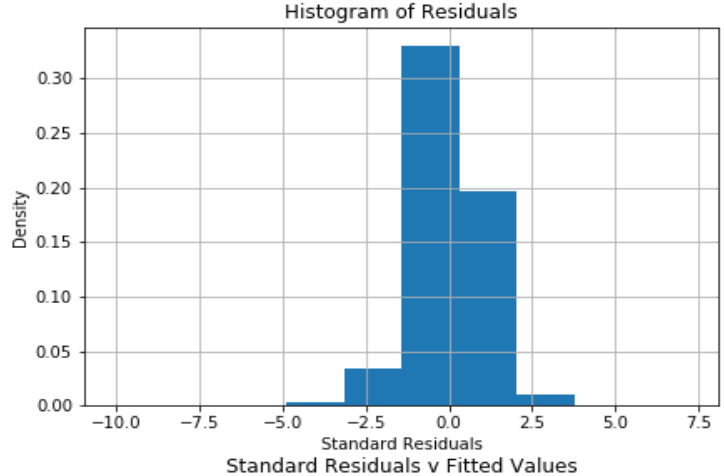

Difference of Random Forest and Out of Bag
by Max Depth of Number of Estimators
Spread of Scores
0.0
0.2
0.4
0.6
0.8
10
20
30
40
50
60
70
Max Depth

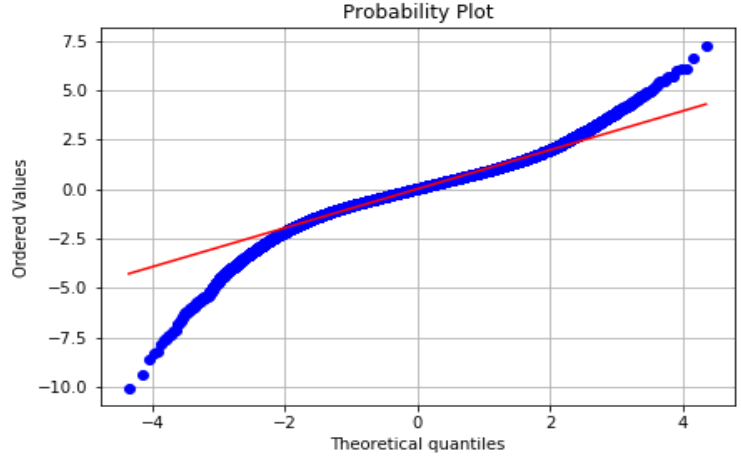
MODEL DIAGNOSITICS ON MONTHS OF SURVIVAL

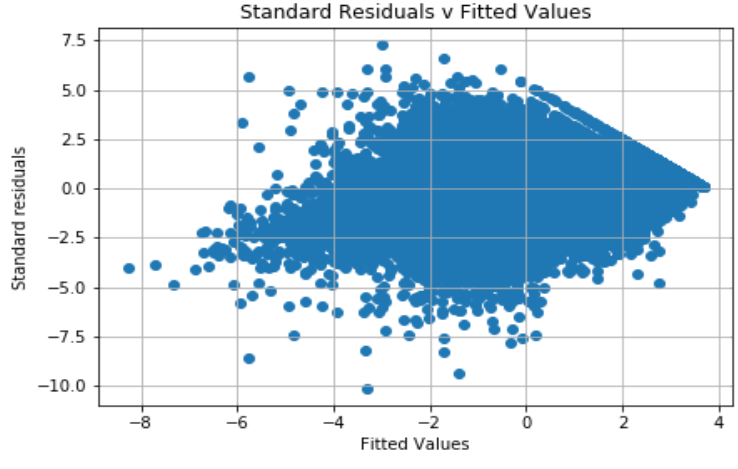
RANDOM FOREST - REDUCED OVERFITTING

Data Set	R^2	Residuals sd	Relative Residual Error
Training	0.31	1.44	0.83
Out of bag	0.31		
Training	0.31	1.52	0.83

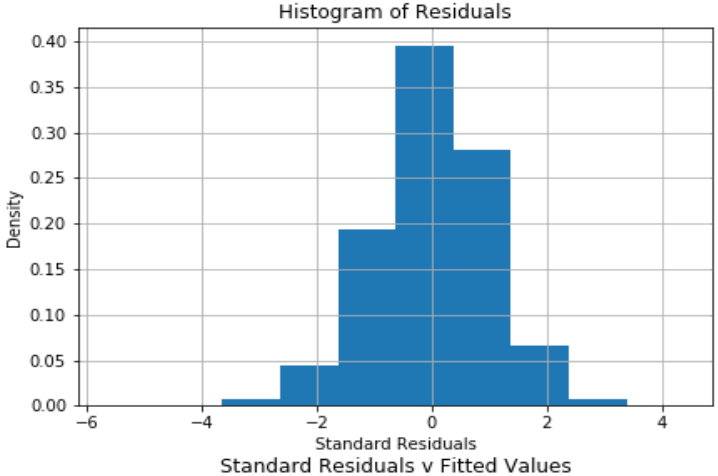
Residual Analysis
Random Forest on Log Odds

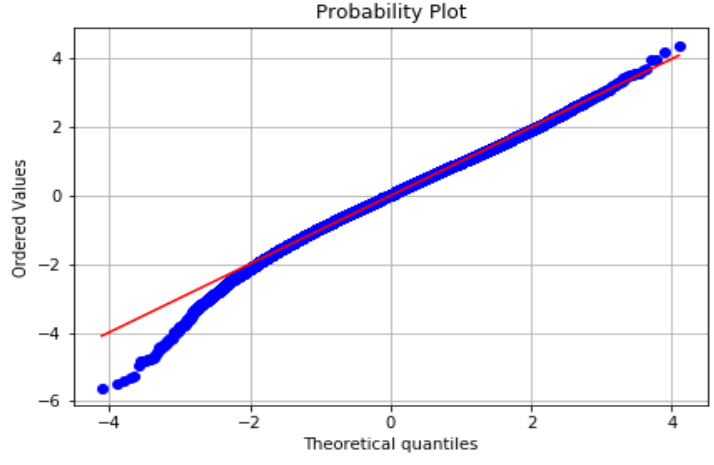


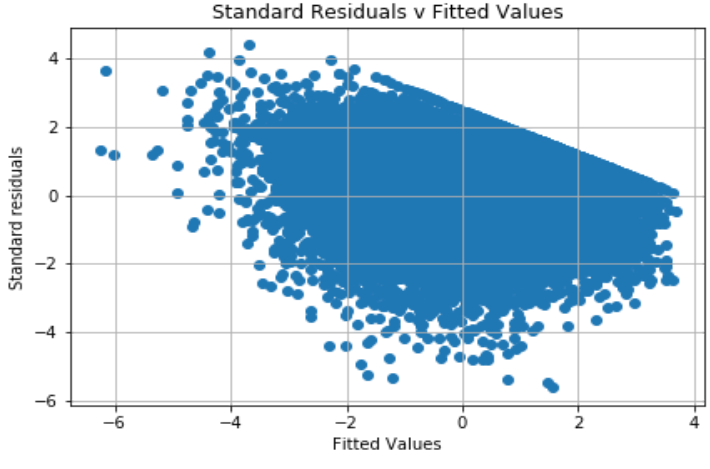




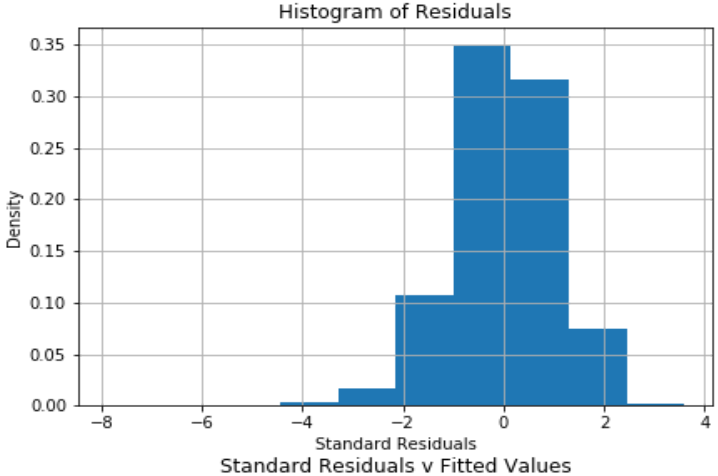
Residual Analysis
Random Forest on Log Odds - Test Set

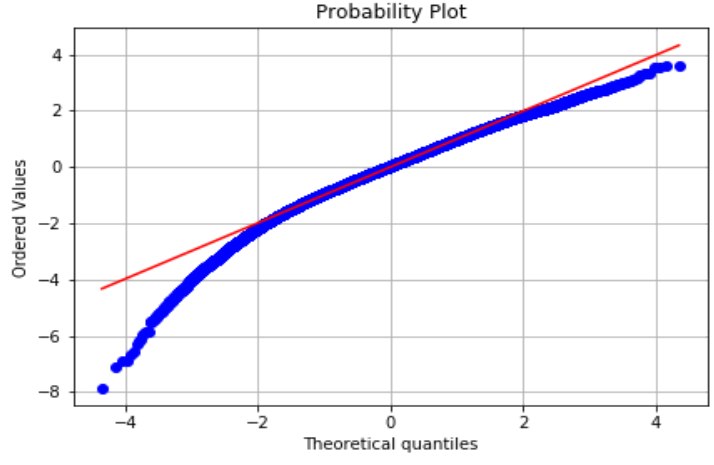


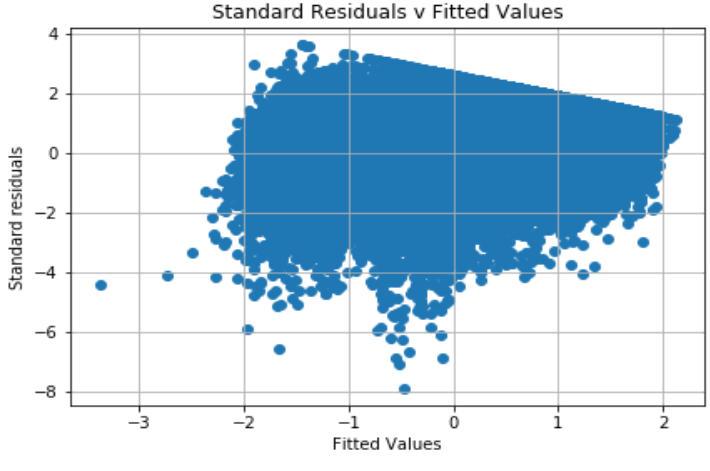




Residual Analysis with Reduce Over Fitting
Random Forest on Log Odds







Residual Analysis
Random Forest on Log Odds - Test Set