GAS PRODUCTION FORECAST

DATA GATHERING

Project Objective:

The purpose of this project is to predict Monthly Gas Production from production gas well characteristics and well treatment operation.

Data Type: Multivariate

Abstract:

Petroleum and gas production are the most important aspect of petroleum engineering. A large capital investment is required to produce oil which also incurs significant annual operational cost to run and maintain drilled wells. Fluid flow meters are usually installed at the bottom of wells to measure the flow of petroleum fluids being produced which is an indicator of how healthy a well is and a measure of the efficiency a workover job carried out on the well. Fluid flow meters are expensive and increase operation/capital cost. Finding an alternate way to inexpensively measure petroleum fluid flow from wells will help reduce cost.

Sources:

Donor
Prof Obadare Awoleke
Department of Petroleum Engineering
University of Alaska Fairbanks,
e-mail: ooawoleke@alaska.edu
TEL: 979-422-5308

Summary Statistics:

Number of instances (observations): 6690
Number of Attributes: 15
Attribute breakdown: 14 quantitative input variables, and 1 quantitative output variable

Variable Information:

Provided is a brief summary of the data context. The order of this listing corresponds to the order of numerals along the rows of the database.

Variable Name	Data Type	Measurement	Description	Variable Type
Perforation Interval	quantitative	feet		Input Variable
Fracturing fluid volume	quantitative	barrels		Input Variable
Proppant quantity	quantitative	Pound mass		Input Variable

Number of Fracture stages	quantitative			Input Variable
Tubing depth	quantitative	feet		Input Variable
Casing depth	quantitative	feet		Input Variable
Flowing Tubing Pressure	quantitative	Pounds per square inch (psi)		Input Variable
Choke size	quantitative	1/64 inch		Input Variable
Shut in Tubing Head Pressure	quantitative	Pounds per square inch (psi)		Input Variable
Specific gravity of gas	quantitative	dimensionless		Input Variable

Well Type	encoded		Deviated = ‘1’,
			Horizontal = ‘2’,
			Vertical = ‘3’
				Input Variable

Latitude	quantitative			Input Variable
Longitude	quantitative			Input Variable
Gas production per month	quantitative	Millions standard cubic feet per month		Output Variable
Acid Treatment	encoded		Acid not pumped = ‘1’, Acid pumped = ‘2’	Input Variable

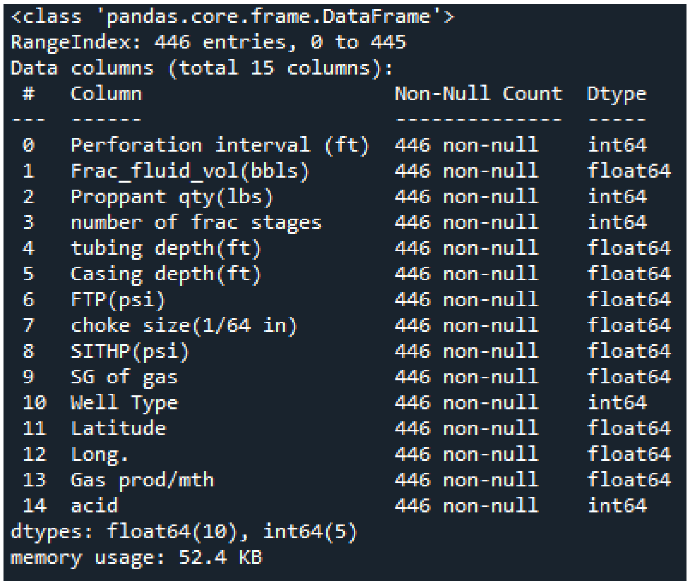
Table 1. Variable information of dataset

Data Characteristics:

The actual gas production rate was gotten from flow meter measurement from an active well. The other variables were measured from same well characteristics and well treatment mixture. Data is in raw form (not scaled). Qualitative “Well type” and “Acid treatment” variables were already pre-encoded from the source.

DATA PRE-PROCESSING

I checked to see if my data had any missing points and it didn't as shown in figure 1.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 446 entries, 0 to 445
Data columns (total 15 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Perforation interval (ft) 446 non-null    int64
 1   Frac_fluid_vol(bbls)     446 non-null    float64
 2   Proppant qty(lbs)        446 non-null    int64
 3   number of frac stages    446 non-null    int64
 4   tubing depth(ft)         446 non-null    float64
 5   Casing depth(ft)         446 non-null    float64
 6   FTP(psi)                 446 non-null    float64
 7   choke size(1/64 in)      446 non-null    float64
 8   SITHP(psi)               446 non-null    float64
 9   SG of gas                446 non-null    float64
 10  Well Type                446 non-null    int64
 11  Latitude                 446 non-null    float64
 12  Long.                    446 non-null    float64
 13  Gas prod/mth             446 non-null    float64
 14  acid                     446 non-null    int64
dtypes: float64(10), int64(5)
memory usage: 52.4 KB
```

Figure 1. Summary of data

All my data type was numerical, hence no need to encode data. I decided to run scatter plots on all my attributes to check for outliers and I found attributes: Gas Production per month, Fracturing fluid volume, longitude and latitude to have outliers which supposedly may be due to human error in data collection.

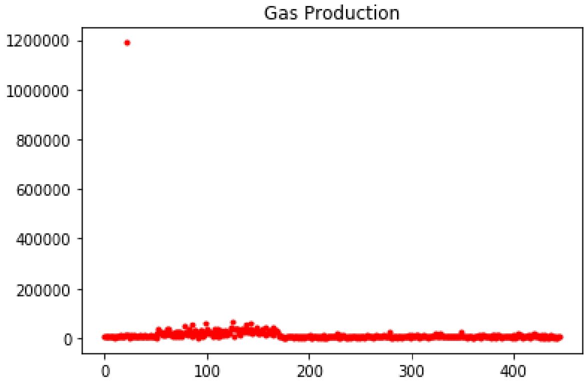


Figure 2a. Plot of gas production per month

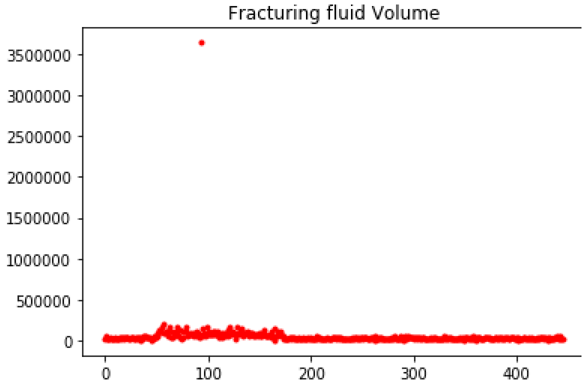


Figure 2b. Plot of fracturing fluid volume

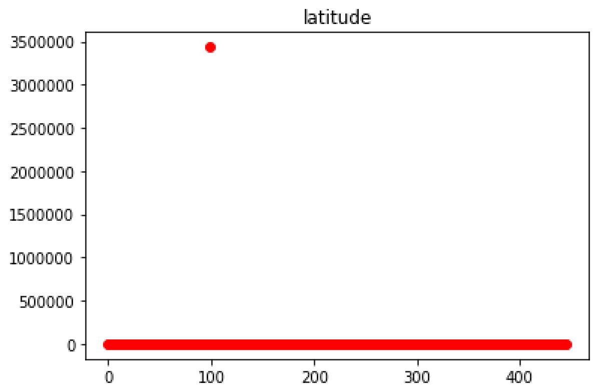


Figure 2c. Plot of Latitude

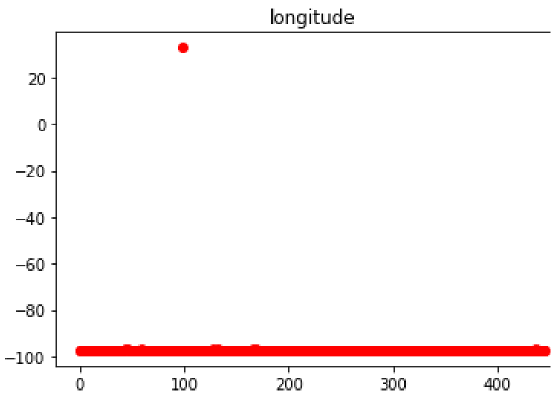


Figure 2d. Plot of longitude

Figure 2 plots above show the outliers observed in these attributes. In other to handle these anomalies, I decided to delete the entire row across all attributes where this anomaly instances were located. This left my data 3 row shorter as two anomalies were observed in the same row.

I plotted a histogram of all my attributes to visualize the distribution of each attribute as shown in figure 3.

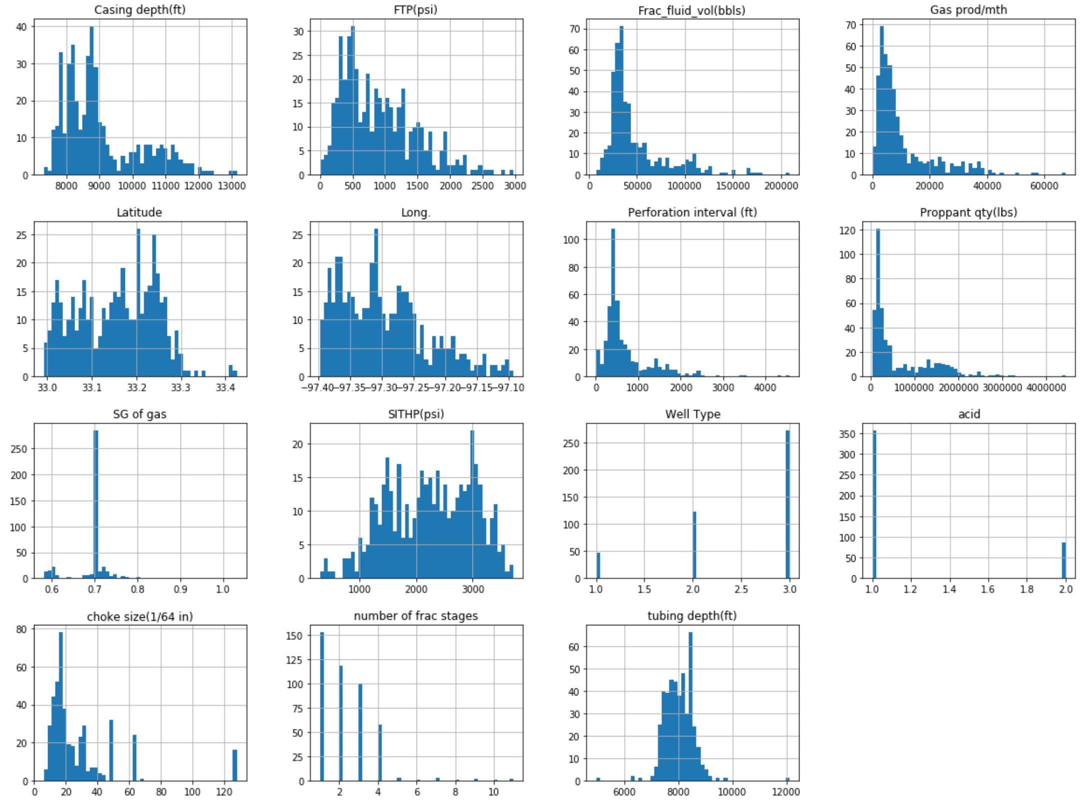
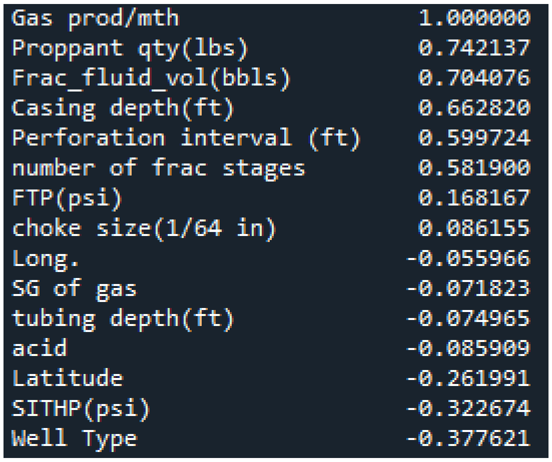


Figure 3. Distribution type of dataset

I observed data attributes to have a highly skewed distribution which will impair the fitness of the model and needed to be standardized. I checked for correlation between predictors and response and also amongst predictors. Linear model will work better when there is linearity between expected value and dependent variable and when there is minimal linear dependence between predictor variables.



```
Gas prod/mth                1.000000
Proppant qty(lbs)           0.742137
Frac_fluid_vol(bbls)        0.704076
Casing depth(ft)            0.662820
Perforation interval (ft)   0.599724
number of frac stages       0.581900
FTP(psi)                    0.168167
choke size(1/64 in)         0.086155
Long.                      -0.055966
SG of gas                  -0.071823
tubing depth(ft)           -0.074965
acid                       -0.085909
Latitude                   -0.261991
SITHP(psi)                 -0.322674
Well Type                  -0.377621
```

Figure 4. Linearity between predictors and response variable

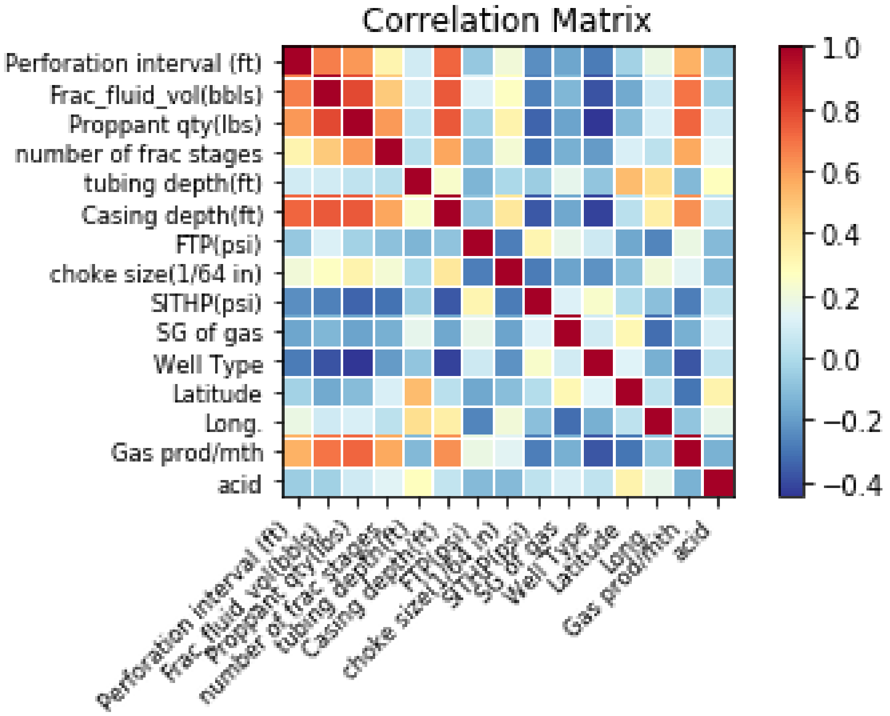


Figure 5. Correlation matrix of attributes

Figure 4 shows the linearity between the predictors and response variable. Half the predictor variables had a linearity strength of at least 0.3, indicating that some of the features will be less significant in defining the response variable. From figure 5 its is observed that there is multicollinearity amongst predictors and some of the predictors will have to be dropped to increase model performance.

I initially divided my dataset randomly into training and testing sets with the 80/20 percentage ratio. I choose to randomly split my data remove any form of bias that may come with splitting the data sequentially. I normalized all the attributes of the training using the MinMaxsScaler approach between 0 and 1 to optimize my model learning algorithm.

Principal Component Analysis and Overfitting Reduction

From the visualization above it was clear that some predictors were more significant than the other and needed to be selected. Also, to reduce model overfitting due to bias I decided to perform a principal component analysis to determine the optimal subset size and features that best describe the response variable. I made use of the linear regression (least square fit) Forward Stepwise Selection technique for this purpose.

Forward stepwise selection begins with a model containing no predictors and then adds predictors one at a time until all the predictors are in the model. I implemented this algorithm on my training dataset and evaluated and validated the best subset based on its adjusted R^2, Mallow's C$_p$, Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC). Algorithm for the forward stepwise selection is:

Let Mo denote the *null model* which contains no predictors

```
• $For\ k = 1, 2, \ldots, n - 1k = 1, 2, \ldots, n - 1$
    ▪ Consider all $n - k$ models that augment the predictors in $Mk$ with one additional predictor
    ▪ Choose the best among these $n - k$ models, and call it $Mk_{+1}$
• Select the single best model among $M_0, M_1, \ldots, Mn$ using cross validated prediction error, $Cp$, BIC, adjusted R$^2$ or any other method.
```

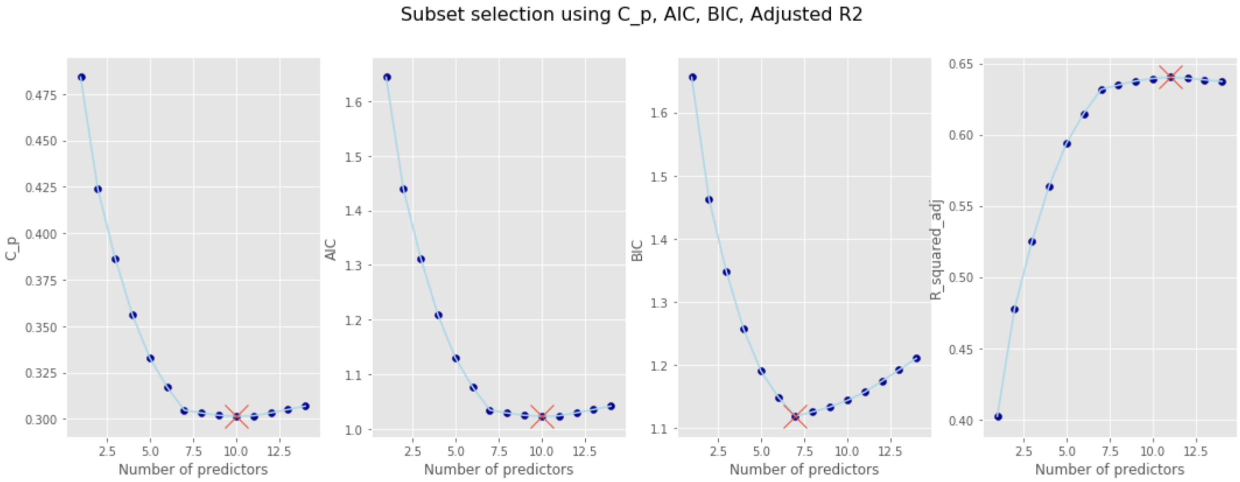


Figure 6. Plot showing optimal subset of each approach indicated by the red cross

Figure 6 shows the result of indirect estimation of test error by making adjustment to the training error using the four different approaches. The feature names for each of this method is listed in table 2. From petroleum engineering stand point, these features should are significant to oil production

Predictors	**Error Standard**	**Subset size**
proppant size, flowing tubing pressure, latitude, longitude, casing depth, well type, choke size	Cp	10
proppant size, flowing tubing pressure, latitude, longitude, casing depth, well type, choke size, number of fracturing stages, perforation interval and SITHP	AIC	10
proppant size, flowing tubing pressure, latitude, longitude, casing depth, well type, choke size, number of fracturing stages, perforation interval and SITHP	BIC	7
proppant size, flowing tubing pressure, latitude, longitude, casing depth, well type, choke size, number of fracturing stages, tubing depth, perforation interval and acid treatment	R-squared	10

Table 2. optimal subset size features

MODELS TRAINING & TESTING

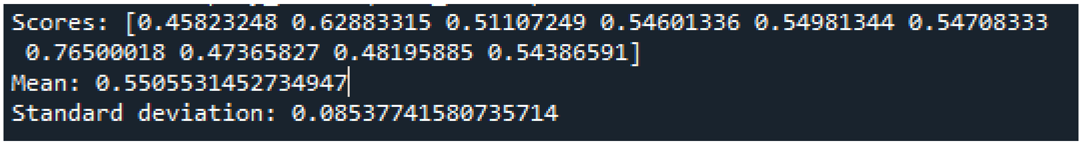
I chose to use linear regression model because there was significant linearity between my predictors and response variable. Also, an adjusted R-square value for the best subset size of 0.63 indicated a decent model fit. I changed my initial data split size to 70/30 which resulted in worse fit based on the error values of those four methods, therefore, I decided to stick with 80/20 ratio.

I ran three different linear regression (least square) different subset size recommended in Figure 6.

Method	Best Value	Root mean square error	Mean absolute percentage error
Adjusted R-squared	0.63	0.53	0.48
Cp & AIC	1.02	0.53	0.48
BIC	1.12	0.54	0.46

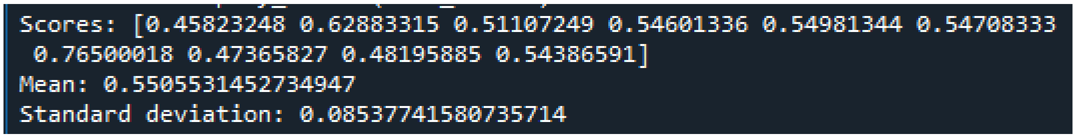
Table 3. Training RMSE and testing MAPE for each of the method's best subset

I cross-validated my model by splitting my training dataset into 10 folds, picking a different training fold for evaluation and training on the 9 remaining folds for each of the 10 runs I made. The results show that Feature suggestion by "BIC" approach had the closet to the mean of the root mean square error calculated from cross validation. Also, the standard deviation was low. This proves that the model is not overfitting the training data.



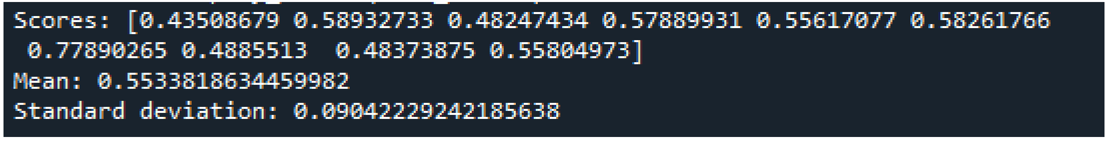
```
Scores: [0.45823248 0.62883315 0.51107249 0.54601336 0.54981344 0.54708333
 0.76500018 0.47365827 0.48195885 0.54386591]
Mean: 0.5505531452734947
Standard deviation: 0.08537741580735714
```

Figure 7a. Cross validation result from adjusted R-squared optimal subset size (11)



```
Scores: [0.45823248 0.62883315 0.51107249 0.54601336 0.54981344 0.54708333
 0.76500018 0.47365827 0.48195885 0.54386591]
Mean: 0.5505531452734947
Standard deviation: 0.08537741580735714
```

Figure 7b. Cross validation result from adjusted Cp & AIC optimal subset size (10)

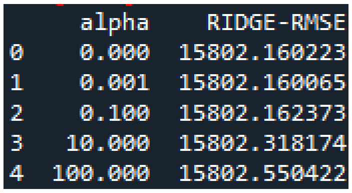


```
Scores: [0.43508679 0.58932733 0.48247434 0.57889931 0.55617077 0.58261766
 0.77890265 0.4885513  0.48373875 0.55804973]
Mean: 0.5533818634459982
Standard deviation: 0.09042229242185638
```

Figure 7c. Cross validation result from adjusted BIC optimal subset size (17)

Therefore, I settled for the optimal subset size of 7 features namely**: proppant size, flowing tubing pressure, latitude, longitude, casing depth, well type, choke size, number of fracturing stages, perforation interval and SITHP**.

I made predictions with my best fitted linear model as above using the test data and made predictions. I the evaluated my model's test error by calculating the mean absolute percentage error (MAPE). The MAPE was **0.46** which is a bit high but still a decent value. I ran ridge regression to try and further minimize overfitting in the model if any by varying regularization strength hyperparameter α to see how this affected my model.



```
     alpha    RIDGE-RMSE
0    0.000  15802.160223
1    0.001  15802.160065
2    0.100  15802.162373
3   10.000  15802.318174
4  100.000  15802.550422
```

Figure 8. Ridge regression result for my test data

The RMSE was pretty the same across all α values indicative of model void of overfitting.

RESULTS

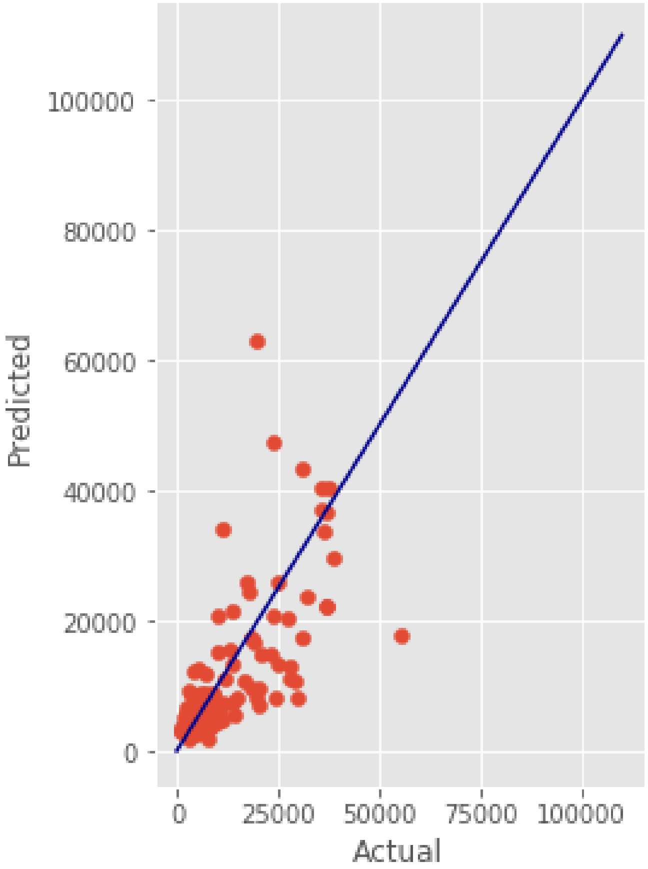


Figure 9. Actual vs predicted plot

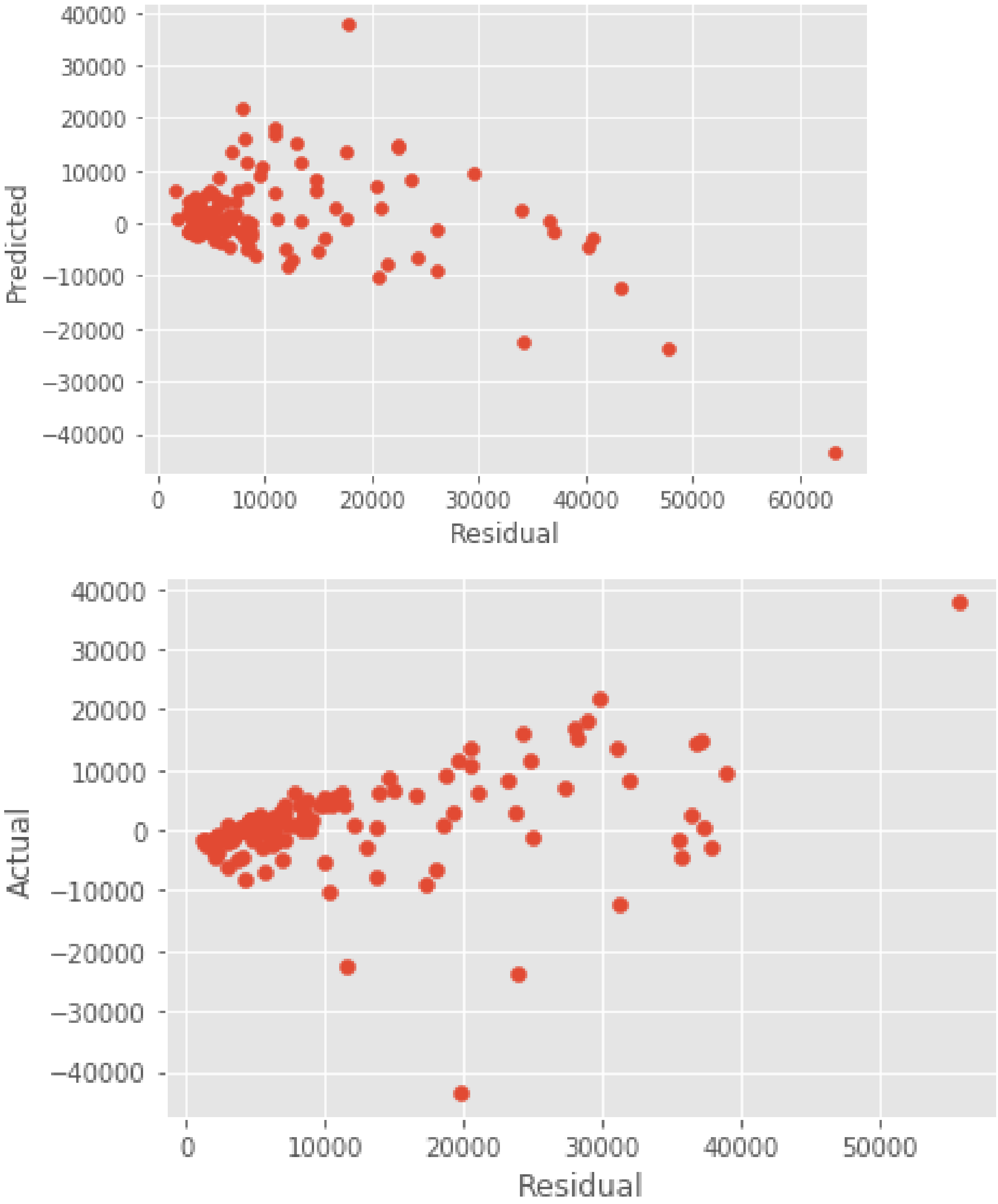


Figure 10. Residual plot

Figure 9 shows a decent fit between the predicted and actual values with small residual errors save some outliers. The residual plot, on the other hand, lacks constant variance as residual error increases with higher response value. This show that the model is doing a fair job in fitting the data.

CONCLUSION

My model fit based on adjusted R-squared value was close to 0.7 which is not bad for a linear model. MAPE for the test set was at a decent high of 0.46. This model is not the best representation of the data. One of the reasons may be that linearity between observations and response variable is not strong true, there may exist some non-linear relation. Another being that more data may be needed to train a better linear model. However, with the cost of flow meters being exorbitant, a 0.46 error in prediction is not such a bad alternative that can be used on wells with same characteristics as those employing the use of flow meter. Predictions can then be adjusted for these wells by using a probability bounds (P10, P50 & P90) calculated by comparing predicted values to measured values from similar wells.