Applying Data Science to Optimize and Maximize Hydrocarbon Recovery from Oil and Gas wells in North America

Contents

Introduction 1
Background 1
Problem 2
Interest 2
Data Description 2
Data Source and Feature Selection 2
Methodology 2
Input Features Histogram and Correlation 2
How does the Bayesian model and Neural Net model work? 4
Results 5
Identifying Significant Drivers – Tornado plots 5
Predicted Intervals based on Probabilities vs Actual Values 6
Neural Net Training and Model Accuracy 7
Discussion 8
Bayesian Model Predictions on the Test Set 8
Neural Net Model Predictions on the Test Set 9
Conclusion 10

Introduction

Background

Uncertainty associated with subsurface modeling leads to a lot of challenges in maximizing oil and gas recovery from wells. The first phase of the shale revolution in north America has led to several trial and errors in the field operations that have lead to either successfully producing from these wells or a complete failure for some of them.

Problem

A huge amount of capital investment (approximately 10 million dollars per well drilled in the US) is associated with the drilling of each oil well and with the current decline in hydrocarbon demand, it has become important more that ever to optimize the drilling of these wells and maximize oil production to supply this source of energy at affordable prices with a sustainable future. With decent amount of data available, it is up to data scientist to leverage this information aptly and apply machine learning techniques to maximize production and reduce the number of dry wells which take a lot of investment but produce no hydrocarbon.

Interest

The machine learning algorithms applied in this project work will be extremely beneficial to oil and gas super major companies such as Shell, Chevron and Exxon. They will also be very helpful to small and mid-cap oil operators to analyze the associated risks and optimize and maximize their hydrocarbon production and hence minimize their economic losses.

Data Description

Data Source and Feature Selection

Texas railroad commission (https://www.rrc.state.tx.us/) has been following a good set of guidelines to follow a record keeping of the location of the wells drilled and some of the key characteristics and properties associated which determine the success or failure of these wells. These properties include but are not limited to the

- Latitude and longitude of the wells
- Length of the wellbore
- Volume of injected fluid and sand
- Injection rates and associated pressures.
- Oil produced

A data pool of 35 wells was collected from this database which has the above mentioned data needed to characterize the clusters of wells that are highly producing so that hidden insights can be unlocked and successfully applied for the drilling of future oil and gas wells.

Methodology

Input Features Histogram and Correlation

Exploratory data analysis consists of plotting histograms of the input features to qualitatively analyze any discrepancies and assure a quick quality control of the input data as seen in Figure 1.

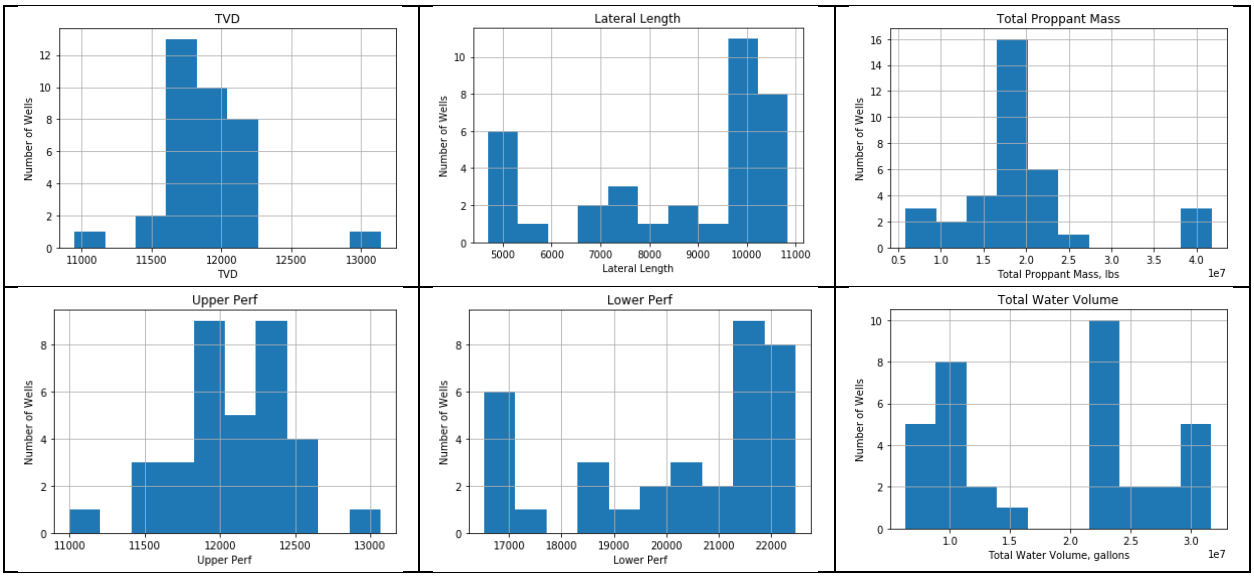
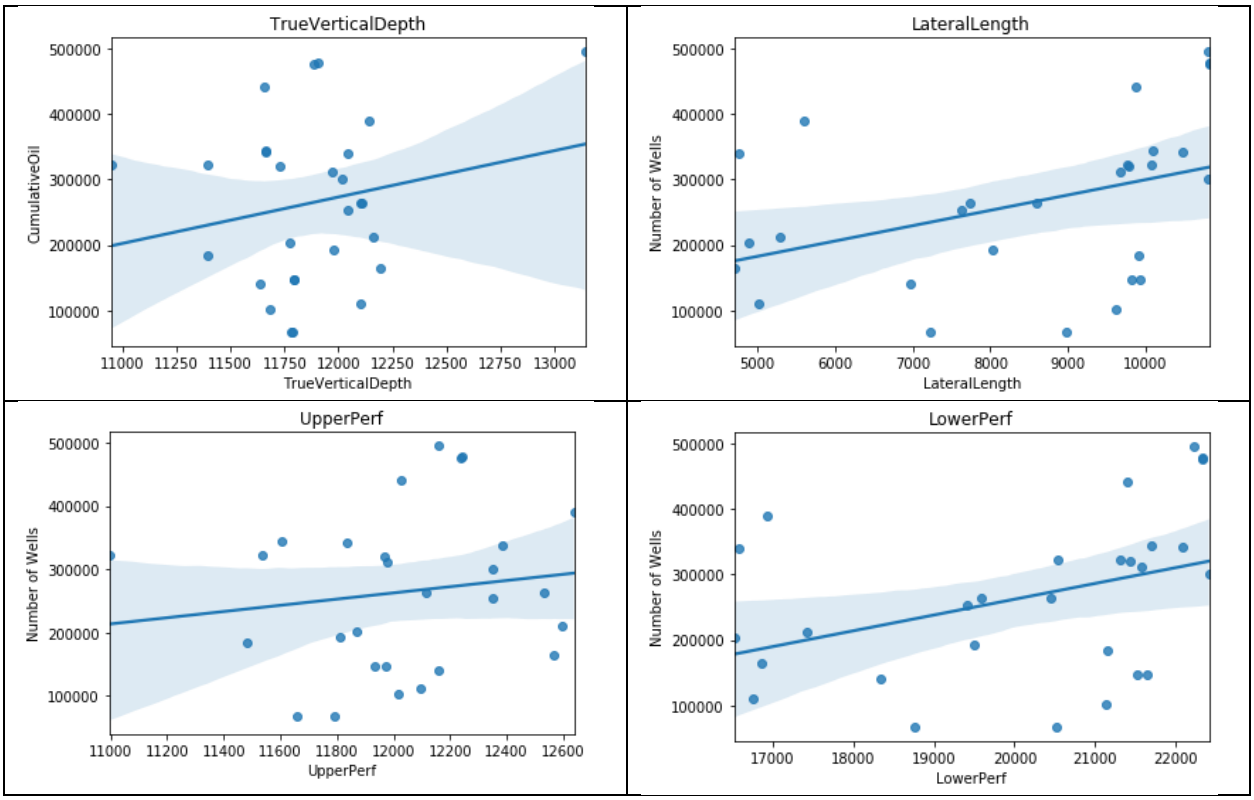


Figure 1: Input Features Histogram

This was followed by descriptive statistics where individual features were cross plot against the target variable which the cumulative oil production per well. Figure 2 shows different positive linear and non linear trends observed between the feature set and the output.



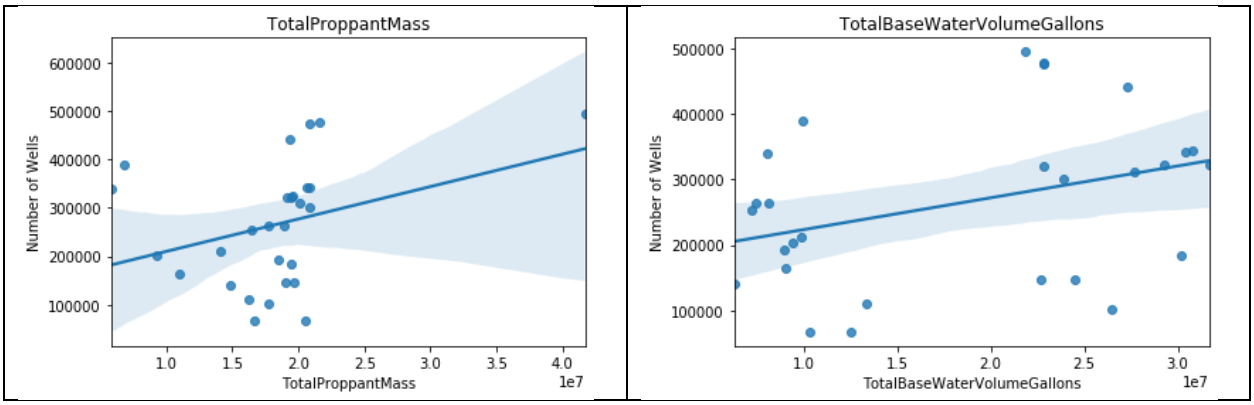


Figure 2: Input Feature vs. Target -- Correlation

How does the Bayesian model and Neural Net model work?

The predictive analytics part mainly consists of 2 main sections

1. Naïve Bayes Classification Algorithm: To analyze and predict probability of well performance (oil production) given input matrix of features.
2. Artificial neural networks: To analyze and predict well performance (oil production).

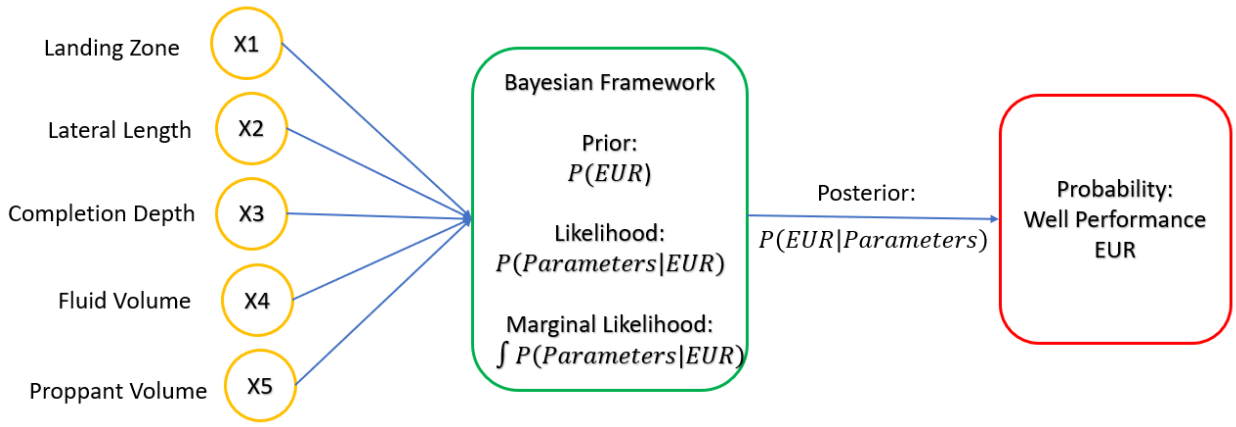


Figure 3: How does the Bayesian Network model work?

Figure 3 and Figure 4 elaborate how the Naïve Bayes classification and the Artificial Neural Network model work, what are their input features used in the analysis, what is the methodology behind and the resulting output in terms of probabilities and the continuous oil production.

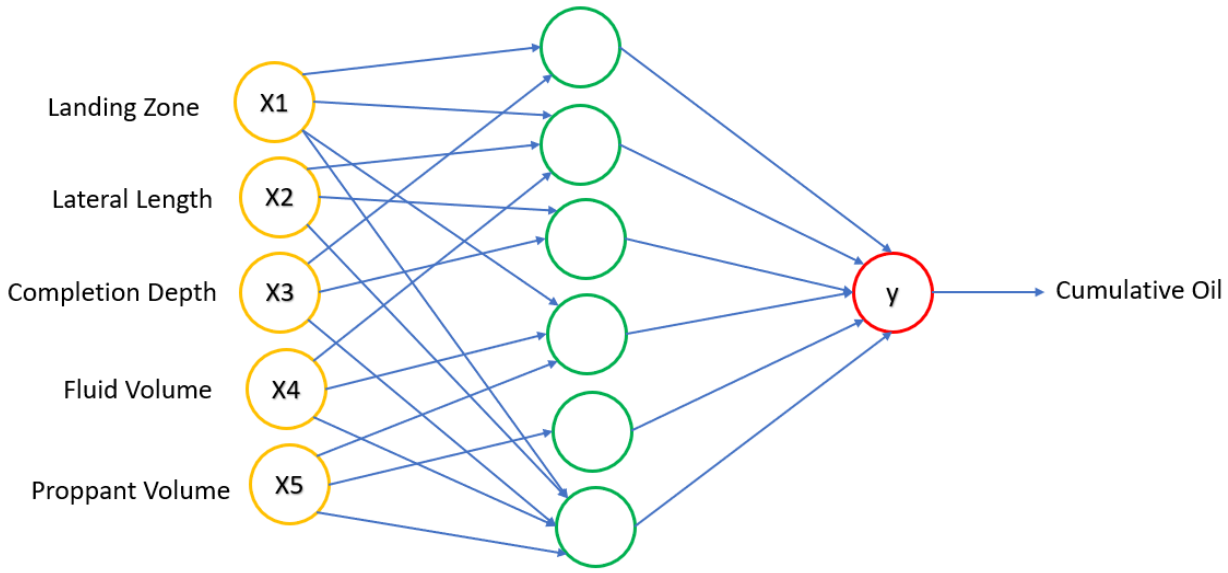


Figure 4: How does the Neural Network model Work?

Results

Identifying Significant Drivers – Tornado plots

Principal Component Analysis (PCA) applied to identify correlations between all the input variables. Based on the identified pattern the number of significant dimensions was combined to 2 principal components (which are now a function of all the significant drivers). The 2 PC dimensions explain more than 80% variance of the input variables. The 2 principal components are now the new hyper-parameters combining all the significant drivers/features into 2 input features. Linear Discriminant Analysis (LDA) gives similar results.

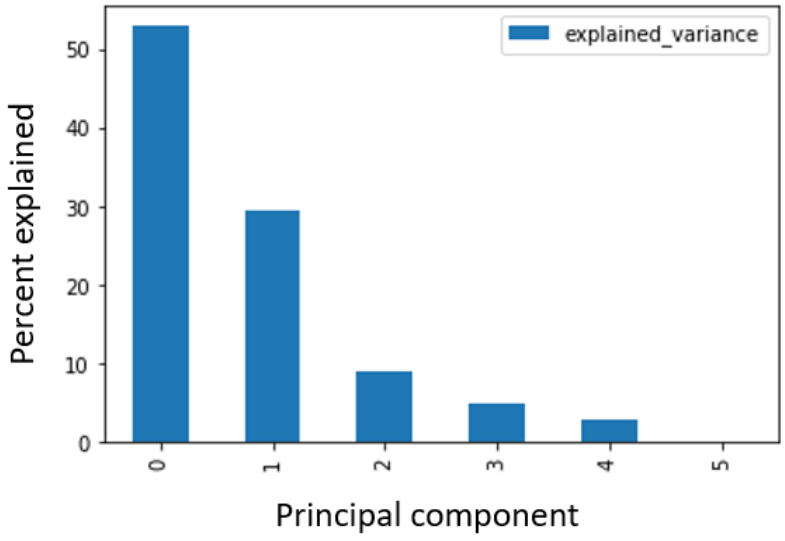


Figure 5: Dimensionality Reduction – PCA

Predicted Intervals based on Probabilities vs Actual Values

Figure 6 shows the map view of predicted intervals based on probabilities from the Naïve Bayes classification algorithm vs. the actual values of the training set. PC1 and PC2 represent the scaled Principal Component Dimensions. The dots represent actual values from wells and the background color represents classifying boundaries and predicted ranges based on the Bayesian model.

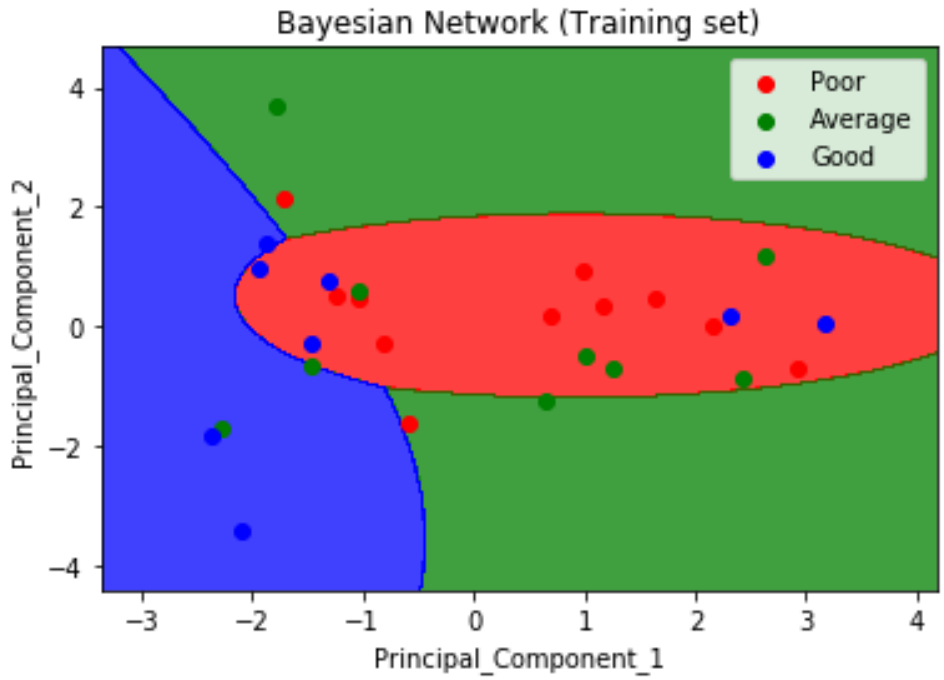


Figure 6: Map View: Predicted Intervals based on Probabilities vs Actual Values – Training Set

Figure 7 shows the actual probabilities associated with the training set predictions and actual values. The confusion matrix shows the training set accuracy of 50%. Dots represent actual values for wells. Area plot shows relative predicted probabilities based on the Bayesian model

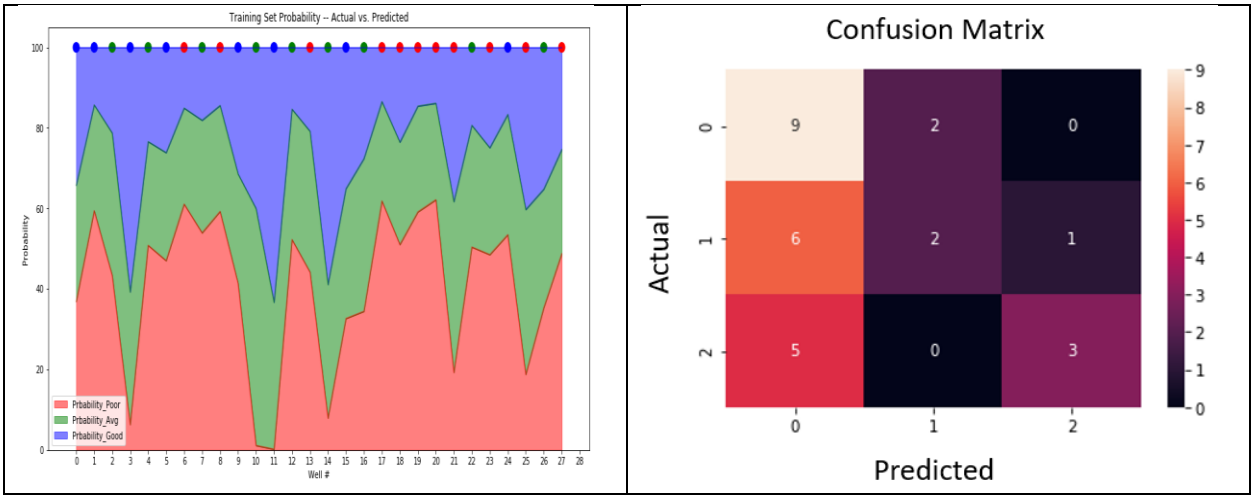


Figure 7: Training Set: Actual vs. Predicted Probabilities

Neural Net Training and Model Accuracy

Figure 8 shows the neural net training and model accuracy as applied to 28 wells. The convergence of the neural net model was achieved after approximately 2300 iterations after which the test loss started increasing which is a sign of over-fitting and the model was stopped at that point. The training set R2 coefficient of 0.449 was achieved.

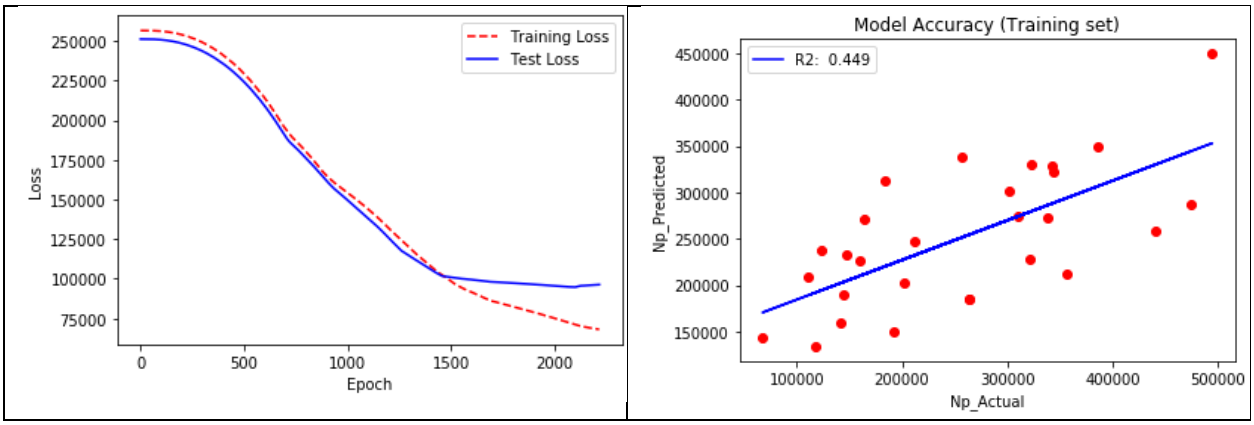


Figure 8: Neural Net Training and Model Accuracy – 28 wells

Figure 9 shows the predicted and the actual values for the produced oil volumes for the training set of 28 wells. As can be seen, the predicted values in the bar chart follows the trends of the actual values quite well. Figure 10 shows their relative errors associated with the predictions.

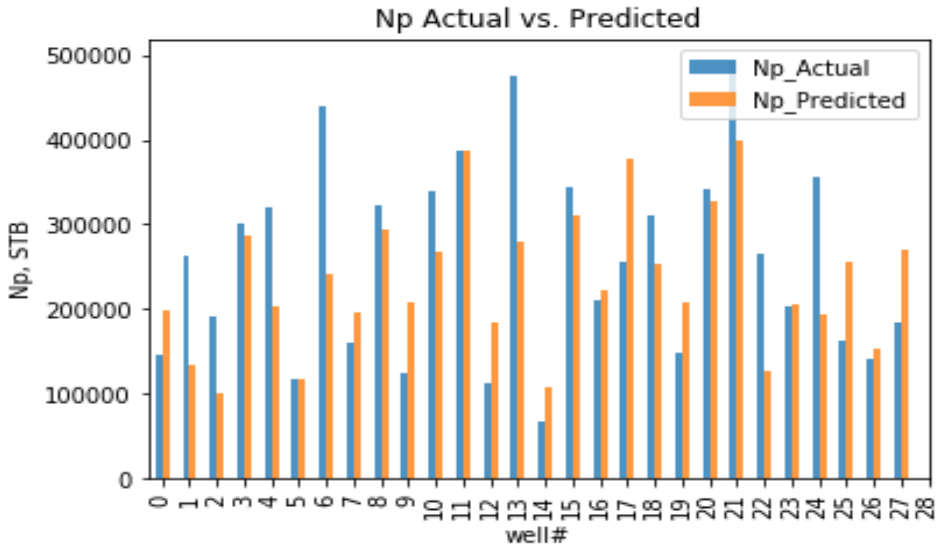


Figure 9: Training Set: Actual vs. Predicted Values

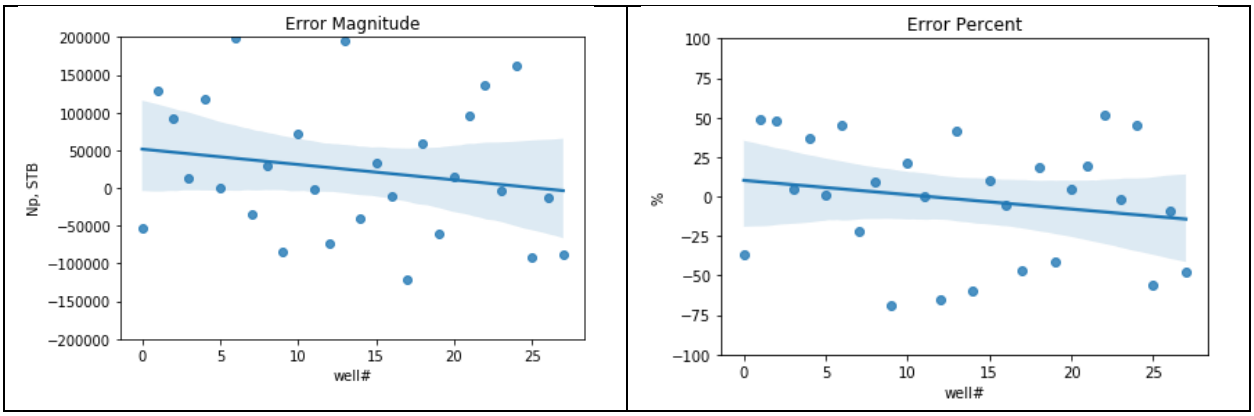


Figure 10: Training Set: Actual vs. Predicted Error

Discussion

Bayesian Model Predictions on the Test Set

Figure 11 shows the map view of predicted intervals based on probabilities from the Naïve Bayes classification algorithm vs. the actual values of the test set. PC1 and PC2 represent the scaled Principal Component Dimensions. The dots represent actual values from wells and the background color represents classifying boundaries and predicted ranges based on the Bayesian model.

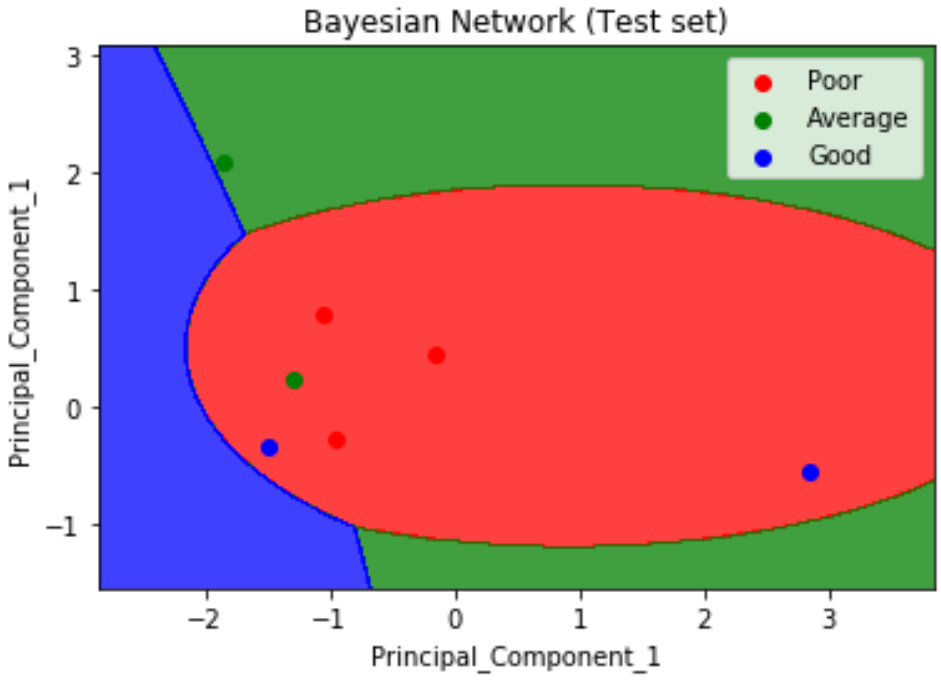


Figure 11: Map View: Predicted Intervals based on Probabilities vs Actual Values – Test Set

Figure 12**Error! Reference source not found.** shows the actual probabilities associated with the test set predictions and actual values. The confusion matrix shows the training set accuracy of 57%. Dots represent actual values for wells. Area plot shows relative predicted probabilities based on the Bayesian model

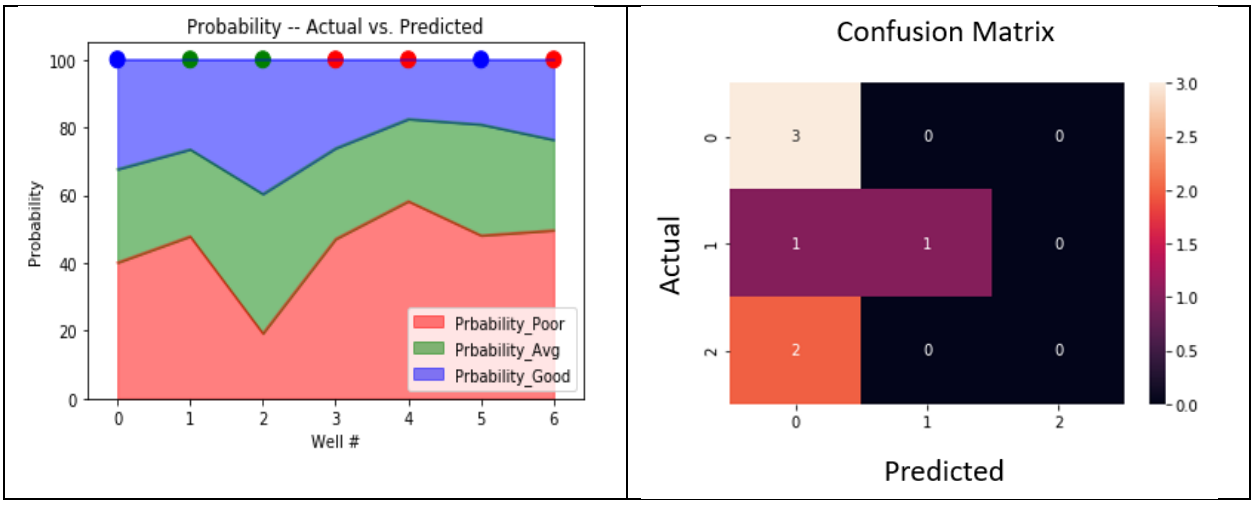


Figure 12: Test Set: Actual vs. Predicted Probabilities

Neural Net Model Predictions on the Test Set

Figure 13 shows the neural net test and model accuracy as applied to 7 wells. The convergence of the neural net model was achieved after approximately 2300 iterations after which the test loss started increasing which is a sign of over-fitting and the model was stopped at that point. The training set R2 coefficient of 0.402 was achieved.

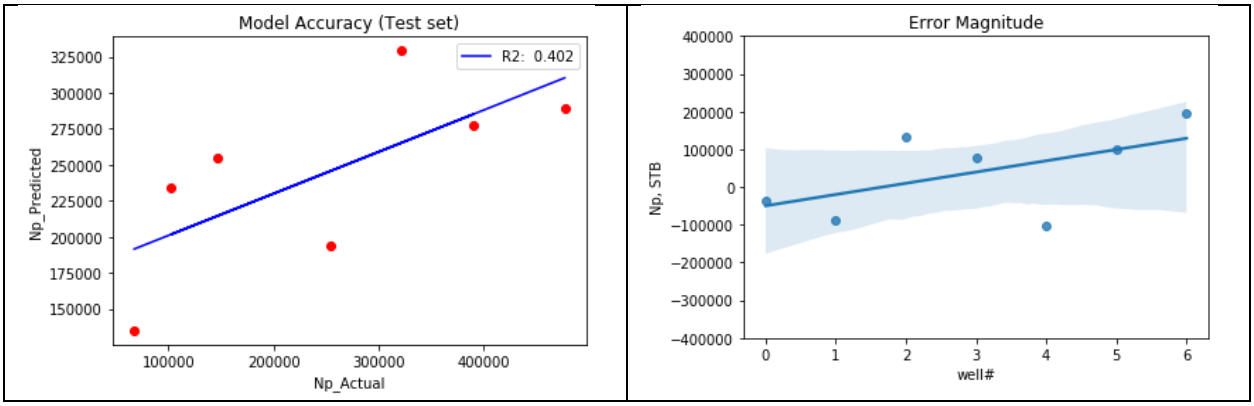


Figure 13: Neural Net Testing Accuracy – 7 wells

Error! Reference source not found. Figure 14 shows the predicted and the actual values for the produced oil volumes for the test set of 7 wells. As can be seen, the predicted values in the area chart follows the trends of the actual values quite well.

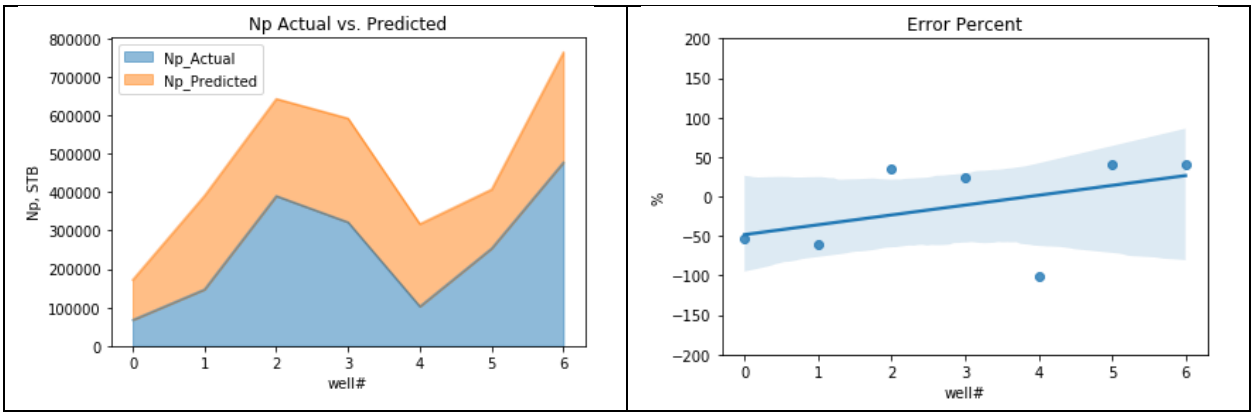


Figure 14: Test Set: Actual vs. Predicted Error

Conclusion

The machine learning algorithms applied in this project work will be extremely beneficial to oil and gas super major companies such as Shell, Chevron and Exxon. They will also be very helpful to small and mid-cap oil operators to analyze the associated risks and optimize and maximize their hydrocarbon production and hence minimize their economic losses.

The models applied in this project classify and predict the oil production from the future wells and what parameters need to be tweaked in order to avoid the potential failure of these oil wells. This will help the oil operating companies adapt their strategy and optimize their field accordingly.