# A Data Driven Approach to Forecast Demand

# A Data Driven Approach to Forecast Demand

Hannah Kosinovsky[1], Sita Daggubati[1], Kumar Ramasundaram[1], and Brent Allen[2]

[1] Master of Science in Data Science, Southern Methodist University, Dallas TX 75275 USA {hkosinovsky,sitad,kramasundaram}@smu.edu
[2] Transformation Office Analytics COE, Flowserve, Irving, TX 75039 {brallen}@flowserve.com

**Abstract.** In this paper, we present a model and methodology for predicting the following quarter's product sales volume. Forecasting product demand for a single supplier is complicated by seasonal demand variation, business cycle impacts, and customer churn. Based upon a Dense Neural Network (DNN) machine learning model, we created a prediction model that considers cyclical demand variations and customer churn. Using the previous five years of parts sales data for a supplier to the oil and gas industry in North America, we found a novel method to predict demand with a minimal error rate [MAE of 0.65]. The Dense Neural Network model performs the best among the other machine learning models we tried in prediction, and additionally, all machine learning models perform better than a non-machine learning solution.

## 1 Introduction

Optimal inventory management is one of the key needs of all manufacturing companies. It is necessary for a company to optimize inventory levels so that their finances are not held up in excess merchandise while making sure that products in high demand are stocked. In financial accounting, inventory turnover is defined as how fast firms sell their inventory items, measured in terms of the rate of movement of goods into and out of the firm [8]. The formula for finding the rate is given by the total cost of goods sold divided by the average inventory within a given time period. Demand forecasting can help a company increase the inventory turns. This means the company is moving the inventory efficiently in the course of business. Ideally, the company will always have the right parts to sell to its customers at the right time; but, without a method to stocking practices, the practice becomes a guess and check method. Overstocking is the result of excess inventory caused by overestimation of demand. Conversely, running out of inventory too fast is caused by underestimating the demand. Having solid demand forecasting is key for efficient inventory management. There are many factors that make demand forecasting complex such as: customer's purchasing behavior, where the products will be used, maintenance requirements of the equipment, and geo-political events that may effect prices beyond previous estimates. The model has to account for as many of these factors that are known to have a prediction that is accurate enough for the company to implement.

In this paper, we present our methodology and model to forecast sales volume for the following quarter for individual products given the previous five years of sales data. Our approach to demand forecasting is to perform machine language techniques like time series ARIMA, Regression Analysis and Survival Analysis. We evaluated the effectiveness of the models using data from an equipment supplier to Oil and Gas industry. We compared these results to the most simplistic forecasting approach used by the industry which is to use the same quarter of the previous years results to forecast the current year quarter numbers. We will call this the "simplistic model".

Given the sales data for five years, we predict the demand such that absolute error of the forecast is minimized. We aggregated the data by quarter and product. Sales volume from the past nineteen quarters was used to predict the following twentieth quarter. Keras dense neural network running on Tensorflow[1] forecasted sales volume by product with a mean absolute error of 0.65. This model performed better than ARIMA, linear regression, elastic net and random forest regression.

Using an ARIMA model, we were not able to forecast future customer demand by the volume, since the data behaved close to white noise. We found that the Dense Neural Network model performed the best among all regressors with the lowest error rate. We found that the anonymized data provided by the sponsor was largely insufficient at providing meaningful explanatory variables.

Expanding on our results of the best demand forecasting solution of Dense Neural Network, we took into account customer behavior often called churn in the industry. Since the sponsor's business is a non-subscription model, it is important to look closely at the purchasing behavior of customers and model those findings in our prediction methodology.

One way we were able to identify customer churn is to perform survival analysis and capture customer lifetime value. This way, we could predict when customers would leave the business (stop purchasing) and also what revenue the company has at stake with each of their customers.

In Section 2, we explain the importance of demand forecasting. In sections 3 and 4 we explain different factors effecting supply planning as well as ways to evaluate customer demand. In section 5 we describe our data and some trends that stood out immediately therein. In sections 6 through 10, we present our methodology and models used for predicting demand. We introduce the metrics we use to evaluate those models in section 11. Our analysis is outlined in section 12. Section 13 outlines potential ethical concerns in this project and how we avoided them. We draw relevant conclusions in section 14.

## 2 Importance of Demand Forecasting

For any company accurate demand forecasting is critical. In industrial companies in particular, not all products are in demand at all times. These companies use

[1] More information may be found at https://www.tensorflow.org/guide/keras

demand forecasting to align production with demand. If a company does not have an accurate forecast, then the company cannot optimize its inventory. One of the many barriers in reaching an optimal model is that there is a constant trade off between ensuring stock of in demand items while also minimizing holding costs. One of the ways this is mitigated is by looking at different demand structures for different items. Another strategy is to analyze the seasonal patterns in which customers order their items. As Nagasawa et al point out, "... If a single ordering policy is applied to several different items that have different shipping statistics, then the inventory will be a mix of items with shortages and excess inventory" [4]. This is why it is important to look at the timing of the orders and also the association between items and their order frequency.

Supply planning is done based on demand forecast. If the actual consumer demand is not the expected demand rate, all the planning done for production, resources, inventory, and distribution will not yield the benefits that are in the business plan and will increase the working capital.

## 3 Equipment Used in Oil and Gas Industry

At a high level, the equipment used in the oil and gas industry can be categorized into critical plant equipment and non-critical plant equipment. The critical equipment is critical to business, safety, and process. This equipment has to perform within the specification to maintain their integrity. Critical and non critical equipment have different demand patterns.

As explained by Sahoo in his book on Process Plant Management, turnaround is a highly-expensive period of regeneration in a plant or refinery [7]. During this period, entire parts of the plant are offline. All critical equipment is inspected and revamped. The demand related to critical equipment is virtually non-existent during this operation of the plant. There is a boom and burst cycle in demand for these parts. During turnaround periods, there is a boom cycle and while the plant is operating the demand goes through burst cycle. Different plants have different turnaround duration. Oil and Gas companies are trying to extend this duration to get maximum use out of their plant equipment. During the operation of a plant, operators identify replacement of major parts or assemblies. Their replacement may require plant shutdown. The parts are ordered around the turnaround time. Equipment manufacturers should have critical equipment and parts required by the oil and gas industry during the burst cycle of the demand. If they do not have the parts, they will lose the business. The forecast model should account for this.

The non-critical equipment have a shorter lifetime compared to critical equipment; since, as their name suggests, the latter is key to the processes in the plant. These will require more frequent maintenance and replacement. Some components that wear over a period of time will need to be replaced at regular intervals to optimize the process. Some of these components that require regular replacements are soft goods like gaskets, diaphragms, and elastomers used for sealing components.

In summary, some parts have seasonal requirements while others have a boom and burst cycle that depends on when a specific company is in a plant shutdown. An optimal forecast model should account for both.

## 4 Customer Churn

For our problem, customer churn is defined as a customer's purchasing behavior changing to reflect a sudden stop or large decrease in orders placed. Customer churn can manifest in different ways for different industries. Companies often spend money to gain customers without paying careful attention to customer retention. This is not optimal, since customer acquisition may cost up to five times more than customer retention as shown in Payne's book on Customer Management [5]. If customers are satisfied and loyal to the brand, it gives the companies potential opportunities to try new products, concepts and ideas without worrying that the customer will leave at the first sign of changes.

In the subscription industry, it is very easy to identify customer churn. If a customer cancels the subscription, it indicates churn instantaneously. For an industrial business, identifying customer churn is challenging. Sometimes, there is simply no demand for certain parts for various reasons like less use overall that slows down the rate of replacement, while other times a customer found the part they needed quicker at another business. There is even a possibility that if another company doesn't have the correct part but is known for quicker service, the customer will simply provide the manufactured part of interest for the new business for them to "reverse engineer". Reverse engineering is a practice by companies that are not the original equipment manufacturer (OEM) to figure out how the part was made by the OEM and then create a similar product for the customer in less time or at a lower cost. The objective of the customer churn analysis is to be proactive and prevent or reduce the flow of customers to other businesses who have more appealing stock or are willing to reverse engineer. The OEM's parts and services revenue is important for overall company performance.

Customer Lifetime Value (CLV) is an important metric in customer relationship management. CLV calculates the future value of a customer over their entire lifetime, which means it takes into account the prospect of a currently inconsistent or low profit customer being more profitable in the future and hence beneficial for the company or organization in the long run [1]. A low CLV can be an indication to the company of potential customer churn. Various firms are relying on CLV to manage and measure their business by capturing the purchase behavior of their customers. In addition, CLV of current and future customers is a good measure of overall value of a firm [6].

## 5 Data

The data that is used in this study is from an equipment supplier having a market capitalization of approximately six billion U.S. dollars. The company has more than 100 well-known brands that are engineered based on customer orders. Some

of the brands have been around for more than 200 years. Their product portfolio caters to oil and gas, petrochemical, power, and other industries. This company's products are at the center of the largest flood control project in Europe as well as at oil fields in the Middle East and Africa. It strives to create extraordinary flow control products.

As shown in Figure 1, the number of transactions per year is steady with reasonable fluctuations from 2014. There was a downturn in the industry from 2015 to 2017. The drop in the number of transactions can be attributed to that. Even though there was a reduction in the number of transactions, it was not significant. Irrespective of market conditions, the need to maintain and repair equipment continues to exist. The year 2014 had approximately 1.5 million transactions, which was the highest in the last five years. The average number of transactions over the last five years is approximately 1.4 million per year. The apparent steep drop off in 2019 is due to partial year data.

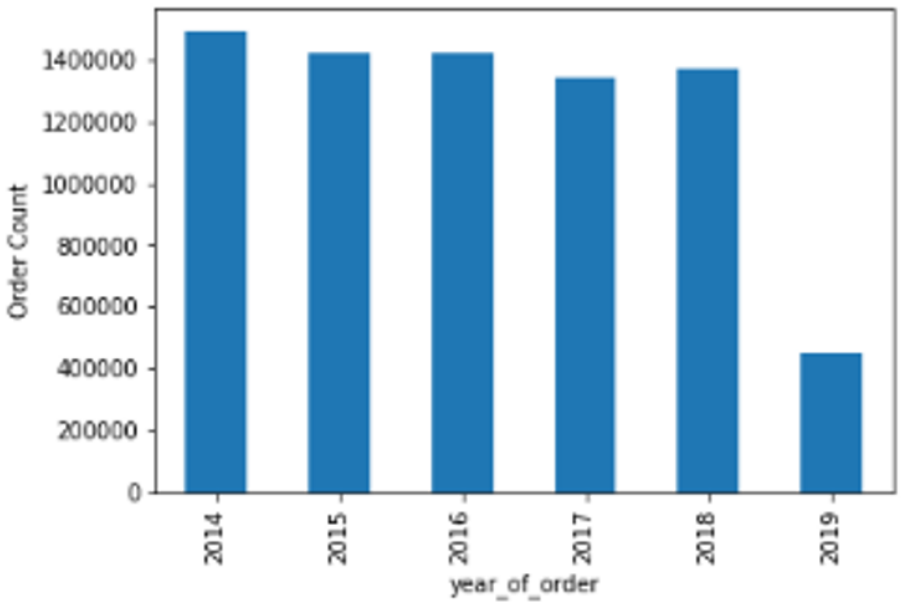


**Fig. 1.** Number of Transactions per Year.

After performing initial exploratory data analysis, we were able to identify important features for this study. The data can be categorized into three groups: order information, product information, and shipping information.

### 5.1 Order Information

The data set contains the following information related to customer's purchase order.

**Table 1.** Order Information

| Feature Name | Description |
|---|---|
| Date | Date of the order |
| OrderNumber | Purchase order number |
| OrderLineNumber | refers to the line item number with in the order |
| Quantity | Order Quantity |
| SellPrice | Sell Price |
| CCN | Customer number |
| PartnerNumber | Partner number |

### 5.2 Product Information

The product information is key data point for company. It is generally maintained in an 'Item Master' inside an Enterprise Resource Planning (ERP) system. There are different consumers for the item master. For example, the supply chain side of business uses information in the item master to purchase parts. The following information in the data set is related to the item and product that is supplied.

**Table 2.** Product Information

| Feature Name | Description |
|---|---|
| Material | material identifier for the part. |
| ItemCategoryGroup | Identifies the item category group. |
| ItemCategory | Item category |
| ItemDescription | Description of the item |
| ProductCode | Product Code |
| ProductDescription | Product description |
| NounCodeDescription | Product grouped into different categories |

### 5.3 Shipping Information

The following attributes are related to where the product is shipping from and the customer's location. This data is helpful if the company wants to stock the products most often bought by a given customer close to that customer's location.

**Table 3.** Shipping Information

| Feature Name | Description |
|---|---|
| ShipFrom | Location where the product is shipping from |
| CustomerName | Customer name |
| CustId | Customer ID |
| City | Customer city |
| State | Customer state |
| PostalCode | Customer zipcode |
| Country | Customer country |
| Region | customer region |
| SubRegion | customer's subregion |
| Industry | customer's industry |
| IndustryGroup | industry group |
| Plant | plant where the product is shipping from |
| $PlantType$ | type of plant |
| $PlantPlatform$ | platform the plant belongs to |

There are nearly a thousand product codes. However, as we see from Figure 2, only 10 product codes contributed to three quarters of total shipments.

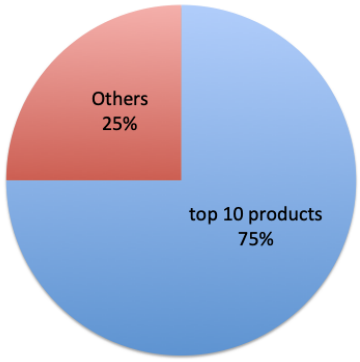


**Fig. 2.** Product Code Shipment Detail

There can be many reasons for some part numbers to have significantly more volume within the data set than others.

The equipment supplier caters to Oil and Gas, Power and other industries. Some product codes cover all industries whereas others may be designed for a specific process in an industry.

Some products may wear out faster than others and the cost of replacement may be cheaper than repair. Customers may decide to replace the parts, driving the volume to increase for these items.

Some of the low volume products can be made of special material that do not wear out as quickly. Customers do not need to replace these items often. Original Equipment Manufacturer (OEMs) may not wish to stock the low volume products and instead opt to make them to order, whereas they will chose to stock high volume products.

Customer shipment data represented in Figure 3 shows a similar pattern. Even though the manufacturer serves thousands of customers every year, only 20 contributed to approximately half of the total shipments.

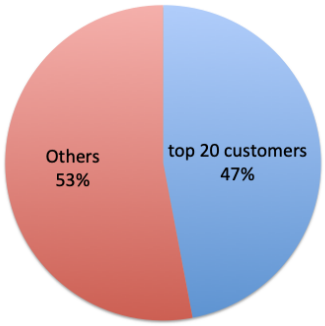


**Fig. 3.** Customer Shipment Detail

In the Oil and Gas industry, there are super-majors, commonly known as "Big Oil" like Shell, Exxon Mobil, BP, Chevron and National Oil Companies like Saudi Aramco, Kuwait Oil Company. These organizations have multiple manufacturing locations and therefore purchase a larger volume of parts than other customers. Their route to market varies based on the geographical region. In North America, some companies use intermediaries to buy products and some buy directly from the manufacturer. These intermediaries effectively function as large clients.

Figure 4 shows the monthly orders for the last five years. At the volume of order level, there is no obvious periodicity visible in the plot. Different customers may have different repair and maintenance schedules. To determine potential periodicity, customer-wise and product-wise analyses are performed.

## 6 ARIMA model

The Autoregressive Integrated Moving Average Model (ARIMA) has three main parts that are represented by a parameter. The autoregressive portion of the model has a "p" factor, the Integrated portion has a "d" factor, and the Moving Average portion has a "q" factor. In this way the model can be expressed as ARIMA (p,d, q). Models without any d factors will be expressed as ARMA(p,q)

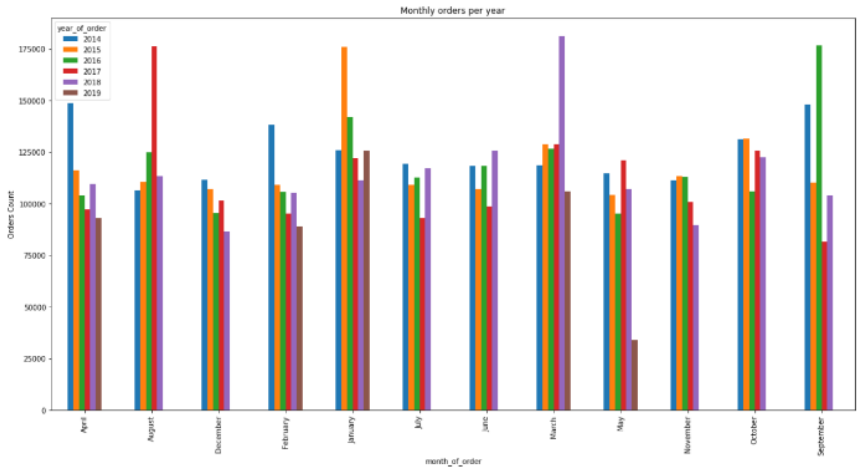


**Fig. 4.** Monthly Orders

models and models that don't have a moving average component will be expressed as AR(p) models, and so forth.

An AR model is one that has a dependent relationship between an observation and some number of lagged observations. It is used to tell the ARIMA model the number of lag observations included in the model. For example, if p =3, it means that we need to use three previous periods of the time series in the autoregressive portion of the calculation [9].

In an ARIMA(0,1,0) model, the value of $X_t$ is equal to $X_{t-1}$ plus a random, zero mean noise component. This means that at each time t, the process is equally likely to move in either direction from time $t-1$. In practice, the d component of the data is found by noting the number of times the data is differenced in order to remove trends - or make the data stationary. If the data was differenced twice, d would be 2.

The Moving Average model is a finite general linear process and is always used to model stationary data. It is not as effective by itself as an Autoregressive model, and is usually used in combination with an AR model. The q parameter gives the size of the moving average window.

## 7 Linear Regression

Regression analysis is a form of predictive modeling. This modeling technique is used for forecasting and finding the causal relationship between continuous variables. The most common form of regression analysis is linear regression and it is widely used for prediction and forecasting. It relies on finding the coefficients of the "best" linear or polynomial fit between a dependent variable and the independent variables – where "best" is determined by minimizing the residuals that

represent the difference between the actual result and the polynomial fit of the regression curve. A residual plot is commonly used to evaluate the regression fit. One of the important indicators that a model is a good fit is if the residual values are equally and randomly spaced around the regression line. Linear regression is a low variance and high bias model and is highly sensitive to outliers.

### 7.1 Elastic Net Regularization

Elastic Net Linear regression Regularization is a hybrid of the Lasso(L1) and Ridge(L2) regularization methods used for linear (or logistic) regression. It is an ideal type of regularization to perform on a model to prevent over-fitting. Lasso regularization reduces the number of features by reducing the sum of absolute values of the coefficients in the model to predict the target variable. Ridge Regularization reduces the magnitude that each feature has on the model by reducing the coefficient value. In essence, the two different penalties in both models work together to build an optimal linear regression model.

## 8 Random Forest Regressor

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting[2] Random Forest is focused on low bias but high variance unlike linear regression. Each tree is grown based on hyper parameters. The following are the key hyper parameters in Random Forest Regressor

$n - estimators$ - specifies the number of trees

$max - depth$ - defines the maximum depth for each tree

$min - samples - split$ - specifies the minimum number of samples required to split a node

$min - samples - leaf$ - specifies the minimum number of samples required in each leaf

## 9 Dense Neural Network

Neural Networks are defined as a set of algorithms that are loosely modeled after the human brain that are designed to recognize certain patterns. As the name suggests, the network layers consist of neurons. Each neuron in a layer receives an input from all the neurons present in the previous layer thus making them densely connected. In the network each input is fed in to a weight that represents the strength of the connection. The weight has the ability to decrease the importance of the input value in the model.

[2] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble. RandomForestRegressor.html

The architecture and some of the key elements of a Neural Network are as follows.

*Activationfunction* - The activation function takes the input and transforms the input relation into an output by adding non-linearity to the inputs. Some common activation functions are Sigmoid, hyperbolic tangent(tanh), rectified linear unit(ReLU) and softmax.

*Epochs* - The Epoch tells us the number of times all the training data have passed through the neural network in the training process.

*AdamOptimizar* - Adam Optimizer is alternate to the classic stochastic gradient descent algorithm that combines Adaptive Gradient Algorithm (AdaGrad) and Root Mean Square propagation (RMSProp).

*BatchSize* - Batch size refers to the number of iterations of training samples used to propagate through the network.

## 10 Survival Analysis

Survival analysis is a type of analysis used to find the expected length of time before an event occurs. It can help identify customer churn if the event we are predicting is the customer no longer making purchases at the company.

The Kaplan-Meier Estimator is commonly used to estimate the survival function from lifetime data. The Kaplan-Meier function is shown as a series of declining horizontal steps which approaches a true survival function for that population.

One of the main advantages of the Kaplan-Meier function is that it can handle right-censored data; meaning that the data is above a certain point but by how much is unknown. In this way, the function accounts for the fact that a customer may have churned if they are not present in the dataset. In order to generate a Kaplan-Meier estimator, at a minimum two pieces of data are required: a status showing whether a customer is churned or not and an indication of their tenure (length of time spent as a customer).

Customer Lifetime Value(CLV) is a dis-aggregate metric that can be used to find customers who can be profitable in the future. Understanding and acting on customer lifetime value (CLV) is very valuable in handling customer churn since companies can focus closely on high value customers.

## 11 Error Metric and Performance Evaluation

The more accurate the prediction is, the more the company can benefit from the forecasting model, especially for the decision-making process within the business. The accuracy of forecasts is determined by considering how well the model performs on new data that were not used when fitting the model [9].

There are many ways of measuring a model's accuracy. We use mean absolute error(MAE), mean square error(MSE) and root mean square error(RMSE) for determining accuracy. The underlying measure for these three metrics is error

rate. Error rate is calculated by subtracting predicted value from actual value. If the error is small, it means the forecasted volume is closer to actual volume in the test data set, meaning the data set that was used to fit the model.

The MSE measures the average square difference between the actual values in the data and the predicted values.

$$MSE = \frac{1}{n}\Sigma_{j=1}^{n}(y_j - \hat{y_j})^2 \tag{1}$$

The MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

$$MAE = \frac{1}{n}\Sigma_{j=1}^{n}\mid y_j - \hat{y_j} \mid \tag{2}$$

The RMSE is a quadratic scoring rule that also measures the average magnitude of the error. It's the square root of the average of squared differences between a prediction and the actual observation.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{j=1}^{n}(y_j - \hat{y_j})^2} \tag{3}$$

The coefficient of determination($R^2$-score) is the proportion of the variance in the dependent variable that is predictable from the independent variable(s). It provides an indication of goodness of fit and therefore a measure of how well unseen samples are likely to be predicted by the model, through the proportion of explained variance. The highest possible value is 1 and can also be negative meaning the model performing worse.[3]. If the $R^2$ value is close to one, the actual values are closer to the regression line. Which means the forecasted volume is closer to the actual volume in the test data set.

$$R^2 = (VarianceExplainedbythemodel)/(TotalVariance) \tag{4}$$

## 12 Results and Analysis

We were unable to fit an ARIMA model to the data due to the data's similarity to white noise. After running diagnostic plots and tests on the quantity data aggregated by quarter, we see no evidence of AR or MA components.

The overfit table is a method used to determine if lag exists in the data or if a periodic component exists. Based on Table 4, the unit root of the 1-B factor present is not close enough to 1 to be evidence of a lag. This was confirmed with a dickey fuller test on transformed data we ran that differenced the values once. Since the p-value was less than .05, we reject the null in favor of the alternative hypothesis that the series is already stationary without the removal of a lag.

[3] https://scikit-learn.org/stable/modules/modelevaluation.htmlregression-metrics

In addition, the overfit table helps confirm that there is no periodic component to the data. In a purely quarterly seasonal model there is a $1 - B$ factor with a system frequency of 0, a $1 + B$ factor with a system frequency of .5 and a $1 + B^2$ factor with a system freq of .25 The table appears to show weak evidence for periodicity for a quarterly series as well as other periodic series like daily and monthly.

**Table 4.** Overfit table

| Factor | Roots | Abs Recip | System Freq |
|---|---|---|---|
| $1 + 1.4647B + 0.5869B^2$ | $-1.2479 + -0.3831i$ | 0.7661 | 0.4526 |
| $1 + 0.4587B + 0.5843B^2$ | $-0.3926 + -1.2480i$ | 0.7644 | 0.2985 |
| $1 - 0.4497B + 0.5421B^2$ | $0.4148 + -1.2933i$ | 0.7362 | 0.2006 |
| $1 - 1.4463B + 0.5321B^2$ | $1.3591 + -0.1796i$ | 0.7294 | 0.0209 |

Lastly, we ran the aic5.wge() function from the tswge() package in R to identify the best ARIMA model fit test using the Akaike information criterion (AIC) as the criteria of fit, and the test found that ARMA(0,0), has the lowest AIC, confirming that our realization is close to white noise.

Since ARIMA was not a valid model, we fit more complex regression models in order to find a prediction.

We prepared our data by creating a matrix as shown in Table 5. The size of the matrix was 70 x 22. Each row represented a product code and each column represented volume in the quarter. The quarters range from the first quarter of 2014 to the second quarter of 2019.

The data set was split into training and validation data sets. We did not consider 2019 data for modeling. For training, product volume of the first eighteen quarters was used to forecast the volume in the nineteenth quarter.

**Table 5.** Matrix Format for Regression Analysis

| Product Code | 2014Q1 | 2014Q2 | 2014Q3 | 2014Q4 | ... | ... | ... | 2018Q3 | 2018Q4 |
|---|---|---|---|---|---|---|---|---|---|
| product 1 | 12567 | 98677 | 123344 | 98765 | ... | ... | ... | 41852 | 279054 |
| product 2 | 15080 | 118400 | 148013 | 118518 | ... | ... | ... | 34877 | 232545 |
| product 3 | 18096 | 142080 | 177615 | 142222 | ... | ... | ... | 50223 | 334865 |
| ... | 21716 | 170497 | 213138 | 170666 | ... | ... | ... | 60267 | 401838 |
| ... | 26059 | 204596 | 255766 | 204799 | ... | ... | ... | 72321 | 482205 |
| ... | 31271 | 245515 | 306919 | 245759 | ... | ... | ... | 86785 | 578646 |
| product 68 | 37525 | 294618 | 368303 | 294911 | ... | ... | ... | 104142 | 694376 |
| product 69 | 45030 | 353542 | 441964 | 353893 | ... | ... | ... | 124971 | 833251 |
| product 70 | 54036 | 424250 | 530357 | 424671 | ... | ... | ... | 149965 | 999901 |

Since we have more than one independent variable, We fitted a multiple linear regression model to determine if there is a linear relationship between the

continuous variables as shown in Table 5 that would help to predict the total volume of products sold for the next quarter. The residual plot in Figure 5 for linear regression revealed that the residuals are at wide, varying distances from the regression line; confirming the non-linearity of the data. When a model does not have a good prediction performance and generalization power, there is an over-fitting problem. In order to mitigate this, we ran elastic net linear regression and saw some improvements, but not significant from linear regression. In order to further improve the fit, we executed a random forest regression, as it is known to perform well with non-linear data, thereby reducing bias. Our best hyper parameters from the fitted random forest regressor had n-estimators of five, max depth of ten, minimum sample leaf of one, min-samples-split of two. The last model we ran to predict the demand for the next quarter was a Tensorflow Dense Neural Network(DNN).We used the Keras regressor from the python package scikit learn to run DNN. Our Neural Network architecture resulting in the lowest error had a hyper parameter with input shape of one, that uses optimizer 'adam', batch size of 32, the loss function of mean squared error (MSE) and mean absolute error(MAE), along with the activation function 'ReLU'.

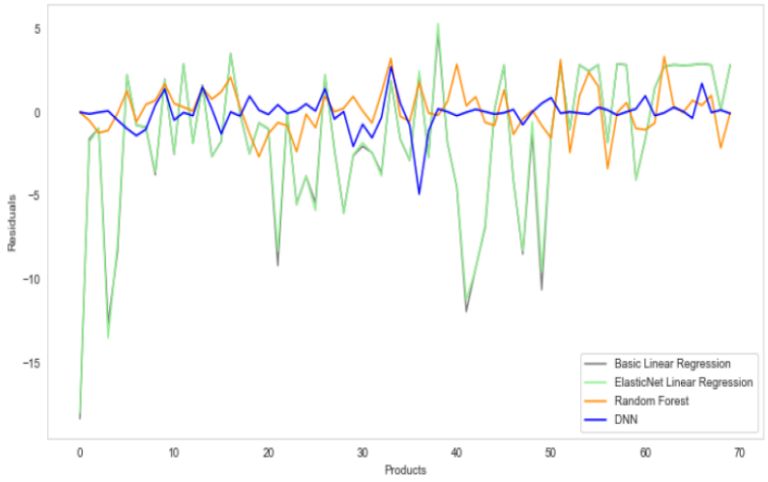


**Fig. 5.** Model Comparison Residual plots

As shown in the residual plot in Figure 5 as well as the bar chart in Figure 6, DNN had the lowest Mean Absolute Error (MAE), Mean Square Error (MSE) and Root Mean Squared Error (RMSE). DNN also had the highest R2 value.

Therefore, out of the multiple machine learning algorithms we used, we found that Tensor flow DNN performed better than the other models for this data set.

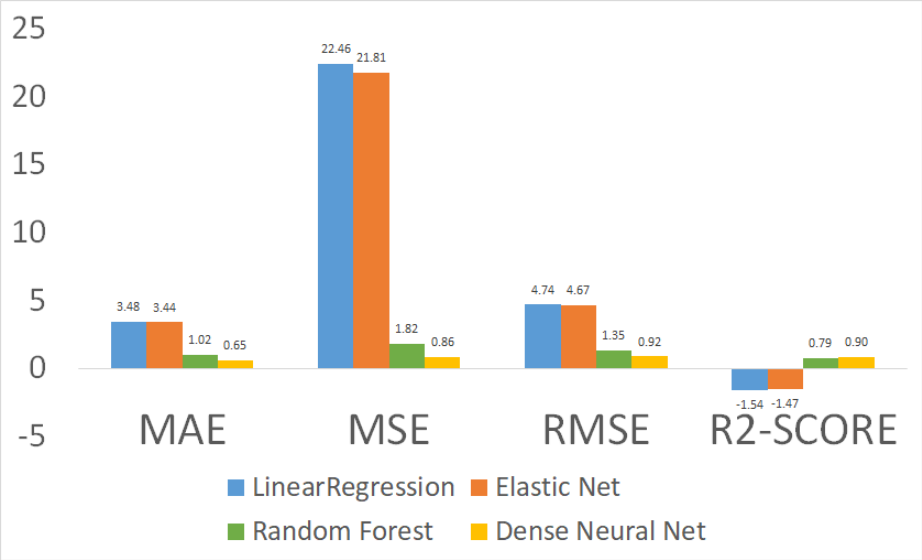


**Fig. 6.** Model Comparison

Table 6 further illustrates the results of each machine learning model as well as the results of the simplistic model. The lowest MAE value of 0.65 was found for Dense Neural Network with R2-score of 0.897.

**Table 6.** Performance Metrics for all Models

| Modelname | MAE | MSE | RMSE | R2-Score |
|---|---|---|---|---|
| Basic Linear Regression | 3.48 | 22.46 | 4.74 | -1.54 |
| ElasticNet Regularization | 3.44 | 21.81 | 4.67 | -1.47 |
| Random Forest Regression | 1.02 | 1.82 | 1.35 | 0.79 |
| Dense NN | 0.65 | 0.86 | 0.92 | 0.90 |
| Simplistic Model | 0.99 | 1.98 | 1.41 | 0.78 |

As mentioned previously, there is no straightforward way to identify customer churn in the dataset since the company doesn't have a subscription based model.

We created a new feature to indicate customer churn. This feature was indicated with 1 if a recent purchase is more than twice the purchase rate, otherwise it was indicated with 0. Purchase rate is defined as the number of purchases over a given period of time.

In order to identify churn, performed survival analysis using the Kaplan-Meier Estimator in the python package lifelines. The purchasing data, which represents all the part orders by customers across the company, was used as the input lifetime data for this estimation method as shown in Table 7.

**Table 7.** Variables used for Kaplan Meier Estimator Models

| count | purchase rate | churned | tenure |
|---|---|---|---|
| 3 | 30.666667 | True | 92.0 |
| 23 | 31.739130 | False | 730.0 |
| 7 | 13.000000 | False | 91.0 |
| 4 | 0.000000 | True | 0.0 |
| 3 | 91.666667 | True | 275.0 |
| 286 | 6.702797 | False | 1917.0 |
| 24 | 57.125000 | True | 1371.0 |
| 1 | 0.000000 | True | 0.0 |
| 2 | 46.000000 | True | 92.0 |
| 17 | 75.176471 | False | 1278.0 |

Figure 7 shows the Kaplan Meier curve for our data. The x-axis represents customer tenure and the y-axis represents the probability of customer churn. Based on the curve, the probability of customer churn of customers having a tenure of 750 days is 45 percent. The customer churn rate does not increase for customers having a tenure between 1375 and 1750 days.

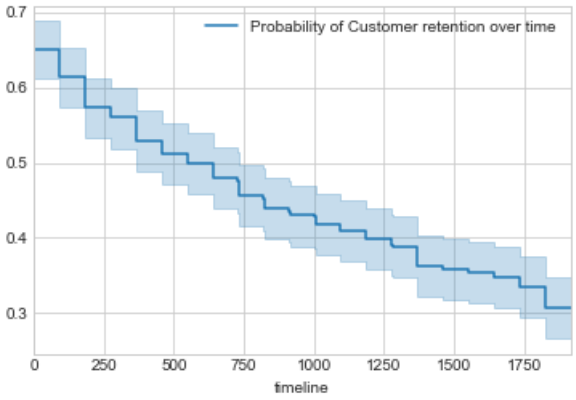


**Fig. 7.** Survival Curve

Figure 8 shows the comparison between predicted tenure and current tenure for customers who have not churned. This gives an opportunity for the manufacturer to take some proactive steps to try to prevent them from leaving.

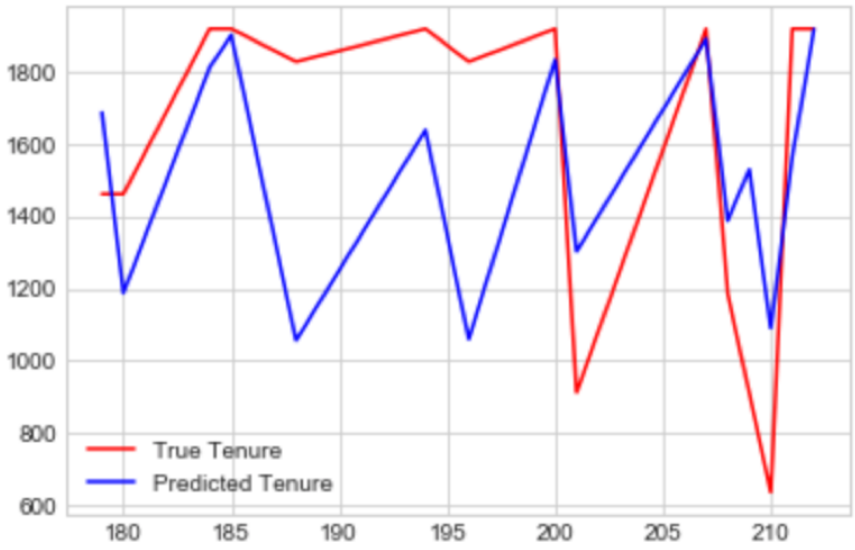


**Fig. 8.** Customer tenure prediction for active customer

To calculate CLV as another method of identifying customer churn, the input data we used consisted of 3 fields: the anonymized Customer Id that made the transaction, the date when the transaction occurred, and the purchase value(quantity multiplied by sale price) of the transaction.

Our data is in transaction form, meaning that there is a separate row for every booking line item and customers can be associated with multiple rows. In order to use this data for predicting the number of purchases made by each customer, we reshaped the data into a Frequency/Recency Matrix using Peter Fader and Bruce Hardie's BG/NBD Model [2]. This means that there will be one row per customer, with columns for frequency (the number of repeat purchases or total purchases minus 1), the age of the customer (time periods since the customer's first purchase), and recency (the age of the customer when they made their most recent purchase)[4].

In order to find the projections for the number of purchases that the customers will make in the next year, we used Fader and Hardie's Gamma-Gamma Model [3] of Monetary Value to calculate the top customers Lifetime Value and

[4] https://towardsdatascience.com/modeling-customer-churn-for-an-e-commerce-business-with-python-874315e688bf

the projected amount spent over the next 12 months. All the models used for calculating CLV are from the python lifetimes package[5].

**Table 8.** Top 5 customers by Projected Total Amount Spent over the Next Year

| CustID | predicted avg purchase | predicted total spent |
|---|---|---|
| cust id 830 | 14638.31 | 1230847.03 |
| cust id 1514 | 5017.12 | 812928.59 |
| cust id 737 | 41585.01 | 554033.91 |
| cust id 3804 | 3273.29 | 541317.71 |
| cust id 207 | 3773.41 | 463805.13 |

The predictions are represented in Table 8 . This information gives the company an opportunity to make informed decisions regarding high value customers.

## 13 Ethics

In our particular case, the only ethical concern is accidental reveal of customer data. We were cognizant to avoid that problem not simply by direct reference to particular customers and their purchase numbers, but by not referencing individual company purchases where the identity of the company could be inferred by the purchase volume.

## 14 Conclusions

As Table 6 illustrates, out of all the machine learning techniques we applied to the problem to forecast demand, only DNN outperformed the simplistic model. Random Forest results approximately equaled that of the simplistic model. Other methodologies we used underperformed the simplistic model. This is an indication that complex machine learning models may be used to better predict forecast demand, while simpler statistical methodologies like regression analysis are unable to predict better than simply using last year's results for the same quarter prediction.