MBAS903: Assessment 2

Hedonic House Price Model & Profit Forecasts

Name: Chinmay Datar

Student No.: 6956361

Table of Contents

Part 1: Hedonic House Price Model

Executive Summary

Real Estate industry is growing rapidly and to keep up with the times it is adapting to new technologies to run the business efficiently. The industry collects data and with recent technological advancements it is rapidly harnessing analytics tools to make the business grow faster. AMES Housing has collected data of 300 houses sold including characteristics of the house and details of sale over the course of 4 years.

Aim of this report is to use this data and create a machine learning model to predict the price of the house. The model utilises the characteristics of the house to predict the price of the house. From our analysis we find that age of the house and area of the house have huge impact on the price of the house when we run the linear regression model. The predicted sales price will not necessarily be the exact figure but will be close to the actual value, suggesting the usefulness of predictive analysis in the real estate industry.

Introduction

In these modern times every industry is trying to incorporate latest technology in day-to-day business tasks to make their work more efficient, save time and maximize profits. Real Estate industry is no different. An industry like Real Estate which plays a key role in every economy it is vital for it to keep up with times and make the process of buying, selling, and renting as efficient as possible to make it more appealing to investors. Australia's real estate has a market size of over $32billion and growing at a rate of 1.5% every year for the past five years. The key factor for the rapid growth of the real estate industry is investors' positive sentiment towards it.

There are many factors that affect the price of a real estate property such as size, shape, location, above ground living area, number of bathrooms, number of car parks, garages, quality of the build, etc. Apart from these factors, other factors, which aren't related to the property itself also influence the price of the property such as supply and demand of the real estate in the country, economy conditions in the country, housing loan interest rates,

demographics of the area, locale, proximity to schools, shopping complexes, etc. All these factors influence the price of a property and willingness of consumers to invest in it. These variables are controllable and can be influenced by the developers, architects, asset managers or estate agents.

The AMES Housing dataset is a collection of 300 houses sold in the time period 2006-2010. The aim of the regression model will be to predict price of any given house based on house features and past property transactions. This will be done by training the model on different subset of features of houses from previous property transactions and testing how efficient it is.

Literature Review

This section is intended to look into previous research done in the field. Looking at journal articles published tackling similar issues will help gain an insight into the current problem. It is imperative to review various models used by researchers in their study.

Chiramel et al. conducted similar research with dataset taken from Ames, IOWA containing 1460 observations (2020, p. 85). The data set contained 79 features in total out of which ones with high degree of collinearity were selected. The variables were OverallQual, GrLivArea, GarageCars, GarageArea, TotalBsmtSF, 1stFlrSF, FullBath. These features were used to train a linear regression model and ridge regularisation model. It was concluded that linear regression model gave satisfactory results with R^2 value of 0.88 on test data.

Another research conducted by Limsombunchai compares using hedonic model and artificial neural network (ANN) to predict price of a real estate in New Zealand (2004, p. 193). The features considered in this study were house size, house age, house type, number of bedrooms, number of bathrooms, number of garages, amenities around the house and geographical location. The result showed that ANN gave better results there were limitations to the findings due to limited data.

Kryvobokov and Wilhelmsson conducted research in Ukraine using hedonic model to predict property prices (2007, p. 157). Their unique approach was assigning weights to the features. The main feature investigated in this study was location of the property i.e., accessibility to

railway station, public transport stops, CBD, water sources and green areas. It also considers prestige of the locale. One of the drawbacks of the study was there were many dummy variables used in order to get the results.

Potrawa and Tetereva took a different approach to the problem and considered online text reviews as well as views from the estate (2022, p. 50). The research also focuses on renting and not transactions of the property. Using image processing the photos collected from online of neighbouring areas and photos of the property it divides the layout and the view into different categories. The other variables were similar to studies conducted previously. Their best model had a R^2 value of 0.74.

From the above studies it is evident that a lot of research has been done in this field. The approaches adopted in many studies is almost similar with many features common to all of them. In the latest study conducted by Potrawa and Tetereva (2022, p. 50) provides a new perspective to the problem and widens the scope for further research in that direction. While Kryvobokov and Wilhemsson's (2007, p. 157) approach to assign weight to features is promising, it lacks the array of features that the other studies have and thus limits the effectiveness of the model. Combining the above studies to create a new hedonic model as well as try to build other machine learning models could achieve more accurate predictions. This can include text analytics for overview of the place, image processing for the property photos and photos of surrounding areas and other variables previously considered in the studies. Our dataset has many similar variables similar to the previous studies conducted and as discussed above. This gives us a direction to conduct our study.

Methodology

The objective of this report is to use the descriptive features of the houses sold and build a regression model to predict price of the houses. We shall use SAS Visual Analytics for visual and statistical representations.

The dataset contains details about the physical features of the house describing age of the house, above ground area, lot size, garage and porch area, condition of the house, material

and finish of the house, etc. These are all important factors which influence the pricing of the property.

We make some assumptions about a buyer's perspective when looking to invest/ purchase a house. We divide these points into four broad categories:

1. The size of the house
2. The condition of the
3. The amenities of the house
4. The aesthetic features of the house

In order to select features for our regression model the features need to be highly correlated with the sales price and must have low or moderate correlation with each other in order to get an efficient regression model. If variables with high correlation are selected for the model the model will take more time to train and will have redundancy.

We shall use the cluster ids created in assessment 1 to use as a classification effect later. We create the first linear regression model using the variables used for the assessment 1 for clustering. Depending on the result of the model we shall add variables which could logically affect the pricing of the house. We can test the model further by adding cluster id as classification effect and check if it improves our result any further. We shall also test interaction effects such as age with condition and quality of the house. The last step to make our model better would be to take natural log of variables which are in square feet as well as take natural log of sales price for the property.

What we are looking in a model is that it should be efficient and shouldn't have redundant variables or variables which logically do not affect the price of a property (low correlation with sales price). To make sure our regression model performs well we divide the data set into training and validation set with 80% of data in training set. To make sure that we get consistent result we assign an arbitrary value to the random state. We shall check if the model should perform equally well on the validation set to make sure the model is not overfit.

Body and Analysis

Correlation Matrix

Before we start making the regression model, we need to create a correlation matrix to see which variables affect the sales price of the house as well as correlation between the variables themselves.

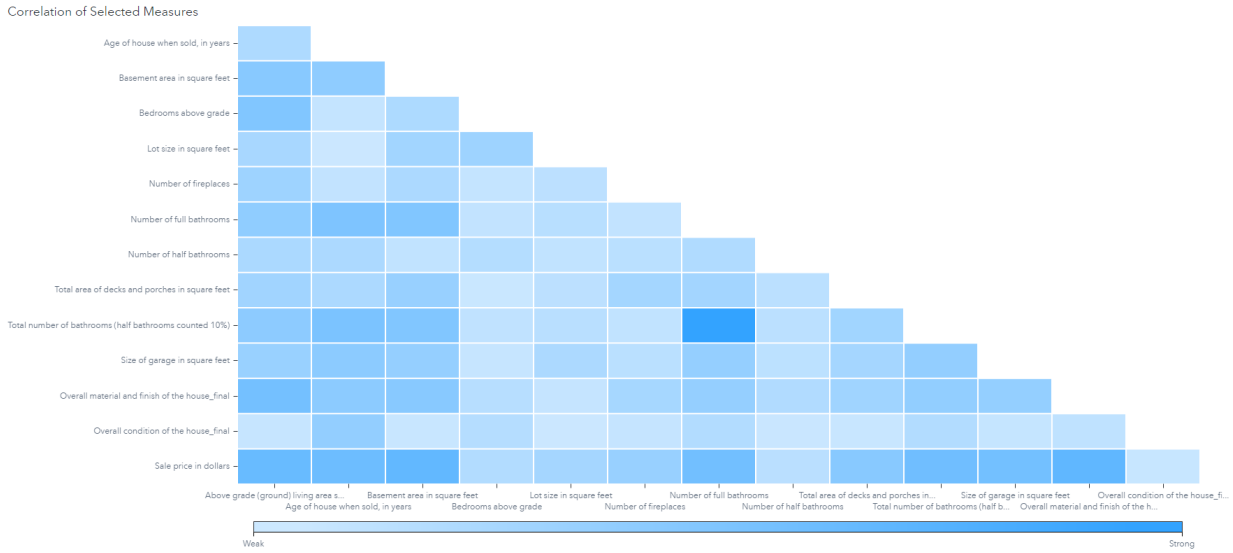

Figure 1 Correlation matrix of all measures

Note: The two columns of 'overall material and finish of the house' and 'overall condition of the house' have been averaged.

Based on this we choose our 6 continuous variables that we shall use for our regression model which have high correlation with sales. The variables chosen are:

1. Age of the house when sold, in years
2. Overall material and finish of the houses_final
3. Above grade living area square feet
4. Number of full bathrooms
5. Basement area in square feet
6. Overall condition of the house_final

This correlates with the four categories that we made in the beginning about buyer's perspective. These variables directly influence a buyer's decision and also the price of the property.

Linear Regression Model

Using the first four variables, we created the clusters. This gave us five clusters or five categories of houses with separate set of characteristics. Using the cluster ids created using the cluster analysis in our regression we find that the cluster id has a p value more than 0.01 so that is not a good variable for the regression model. We also tried to work with interaction effect as mentioned in the methodology, but they were of less importance to the model and had p value more than 0.01. We take the six variables listed above to make a model and got a satisfactory result with a R squared value of 0.8195 (81.95%). To make the model better we take natural log of the variables which are in square feet and also the sales price in order to see how much percentage change in a variable will change the price by one percent.

As can been in figure 2 the linear regression model gives us satisfactory results with R squared value of 0.8777 (87.77%). There are some unused observations because there are 16 missing values in 'basement area' variable. Since we created a partition in the data for training and validation, we can check how well our model performs on unknown data. This will also help us check if the model is underfitted or overfitted in which case the model will not perform accurately.

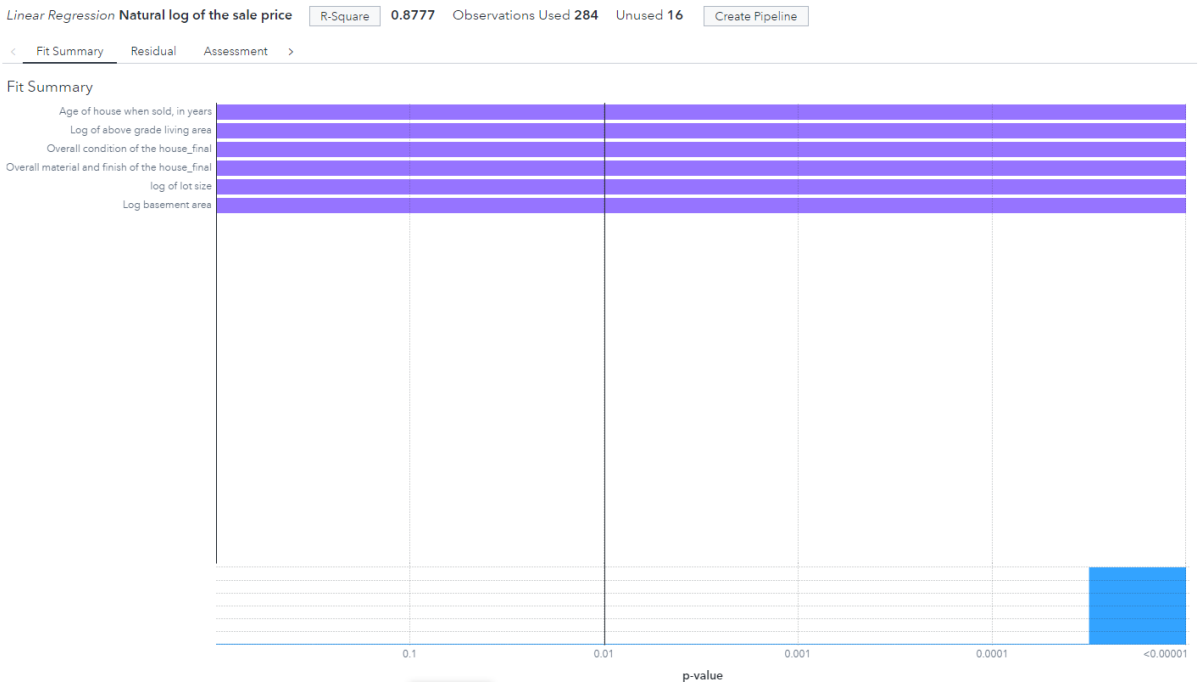

Figure 2 Linear Regression Model

Figure 3 Model Assessment

Figure 3 shows us that model does not perform very well on unknown data which means the model is overfitted on the training data. This is due to the lack of enough observations in the training dataset.

It is important to analyse the residual plot before we state that the model is good or bad. We see in figure 4 that there are few outliers in the dataset. Due to the use of partition, we cannot remove the outliers. Removing a few outliers will help us increase the accuracy of the model but since the dataset only contains 300 observations it won't be recommended to remove them.

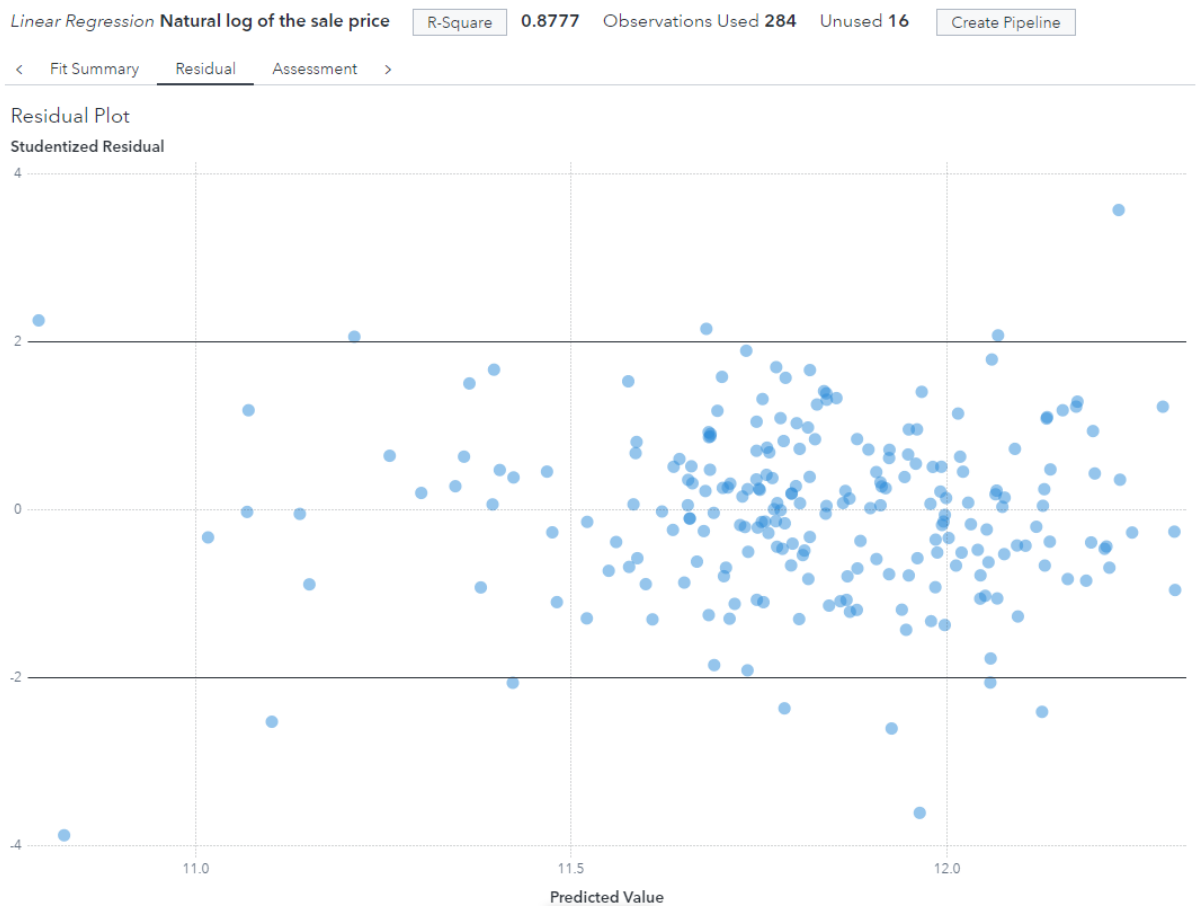

Figure 4 Studentised Residual Plot

The table below makes it easier to understand how change in a variable is affecting the change of sales price. As expected, change in age of house has a negative impact on the sales price. We also see that above grade living area has the maximum impact on the sales price. Increase of 1% in above grade living area increases sales price by 0.39% followed by basement area and lot size. This indicates the increase in any kind of area in the property has a significant impact on the sales price of the property.

Parameter	Estimate
Intercept	6.126
Log of above grade living area	0.3917
Age of the house when sold	-0.0045
Overall material and finish of the house	0.0905
Log of basement area	0.1721

Overall condition of the house	0.0852
Log of lot size	0.1117

Conclusion

The ability to predict prices of the house just by entering the details of the house reduces the burden on real estate agencies to survey the house and come up with a price that can be acceptable to both, buyer, and the seller. Majority of the research done in this field follows a similar pattern with almost same variables but different prediction models. This report takes he same path as other studies. With the right variables both simple and complex prediction models give satisfactory result. By tuning the parameters of the model and gathering data for a greater number of properties the model could give more accurate results.

In our study we notice that factors such as above ground living area, condition of the house and material and finish of the house have a significant impact on the price of the house. Even with some old houses in the dataset the material and condition have high rating which must be due to renovation of the property. This suggests that renovating an old house can increase the price which is beneficial for both buyer and the seller.

The prediction cannot be hundred percent accurate but due to the inherent limitation of nature of the problem. The value an individual is willing to pay for a property depends on his emotions towards the house and reason behind the purchase. The price of a property also depends on accessibility to markets, locale, country's economy, views from the house. It will hard gather all the data and create a simple model including all the variables that influences a buyer's decision. But even with the current amount of research and data gathered the models have high accuracy which gives the real estate agencies a good baseline and then negotiate with the buyer and seller to come to mutually decided value of the property.

References

- Chiramel, S, Logofătu, D, Rawat, J & Andersson, C 2020, 'Efficient Approaches for House Pricing Prediction by Using Hybrid Machine Learning Algorithms', *in Communications in Computer and Information Science*, Springer Singapore, Singapore, pp. 85–94.

- Kryvobokov, M., Wilhelmsson, M.: Analysing location attributes with a hedonic model for apartment prices in Donetsk, Ukraine. Int. J. Strat. Prop. Manag. 11(3), 157–178 (2007).
- Limsombunchai, V.: House price prediction: hedonic price model vs. artificial neural network. Am. J. Appl. Sci. 1(3), 193–201 (2004).
- Potrawa, T & Tetereva, A 2022, 'How much is the view from the window worth? Machine learning-driven hedonic pricing model of the real estate market', Journal of business research, vol. 144, pp. 50–65.

Part 2: Profit Forecasts

Executive Summary

With growing businesses and increasingly complex supply chains it has become a necessity to forecast demands for retailers to keep the business running smoothly and have stock of the appropriate products during peak demand. Product Analysis data set contains sales data collected over a period from 2007 to 2011 with details including order date, costs, retail price, geographical data, supplier details and product details.

In this report we forecast profit for a retailer by product categories and groups to find out which products make the maximum profit for the retailer. We use multiple forecasting models to find out which models suits our use case the best. We use variables which directly affect the profits and consider seasonal trends to get an accurate profit forecast. We also do an 'what if' analysis of where we consider an increase in predicted profits.

Methodology

The objective of this report is to use the transaction details of previous purchases in order to create a forecast for profit. This forecast helps the retailers to stock up and order the right quantity of goods. Lack of accuracy in the forecast can lead to over stock or understock which means loss in business and profit. We shall use SAS model studio to create our forecasting model and assess it.

The data set contains details of products sold by online retailers. It contains details of the supplier, product, quantity, cost, retail price, discount, month of sale and profit along with geographical details of customers such as country, state, and city over the period of 2007 to 2011. We shall use order date as time variable, profit as dependent variable and try different variables for By variable and independent variable.

Next, we shall look at creating pipeline. We will create one pipeline with different forecasting models and then compare the different models. This will include Auto-forecasting model, Naïve model, Hierarchical model, and a Combined Hierarchical model (combined with Unobserved Components model (UCM), Exponential Smoothing model (ESM) and ARIMAX). We will compare all these models and check the weighted mean absolute percentage errors

(WMAPE). We need to change the reconciliation level to see how the results change which is most accurate for our study.

Once the forecasting model is ready, we shall take a look at the time series plot and the forecasting. Secondly, we shall create a hypothetical scenario to check how well our model does when we expect a fluctuation in demand.

Results

Selection of Variables

The variables for forecasting are selected based on at what level the forecasting is expected by the user. Variable 'Order Date' is assigned as time variable for our forecast and 'profit' as the dependent variable. In our forecasting model we are going to forecast profits by product category and Product Group. We select these two variables as By variables. Next, we select the independent variables which will be responsible for the forecasting. For this we select variables which directly affect profit by simple logic. We selected five variables as independent variables as follows:

1. Cost
2. Discount
3. Retail Price
4. Quantity
5. Month

These variables directly impact the percentage of profit made on each product. Variable 'month' will help us predict the correct seasonality of the orders and get more accurate predictions.

We start to create the pipeline for the forecasting model. The auto-forecasting model is created by default. We create 3 more child nodes at the Data node. The 3 new nodes are Hierarchical forecasting node, Combined Forecasting node and Naïve Forecasting node. The final pipeline looks like figure 5.

We checked the reconciliation at each level. We found that the model worked best at reconciliation level of Product_Group.

Figure 5 Forecasting Pipeline

We now check the results of the model comparison. The results of the model were as follows:

Model Name	**WMAPE**	**WMAE**
Auto-forecasting	52.8384	409.7153
Hierarchical forecasting	58.8951	391.6021
Combined forecasting	58.6497	390.4480
Naïve Model	91.4074	588.6277

From the above table we can conclude that Auto-forecasting is the clear winner with lowest WMAPE value of 52.8384. Auto-forecasting model is our champion model in this case.

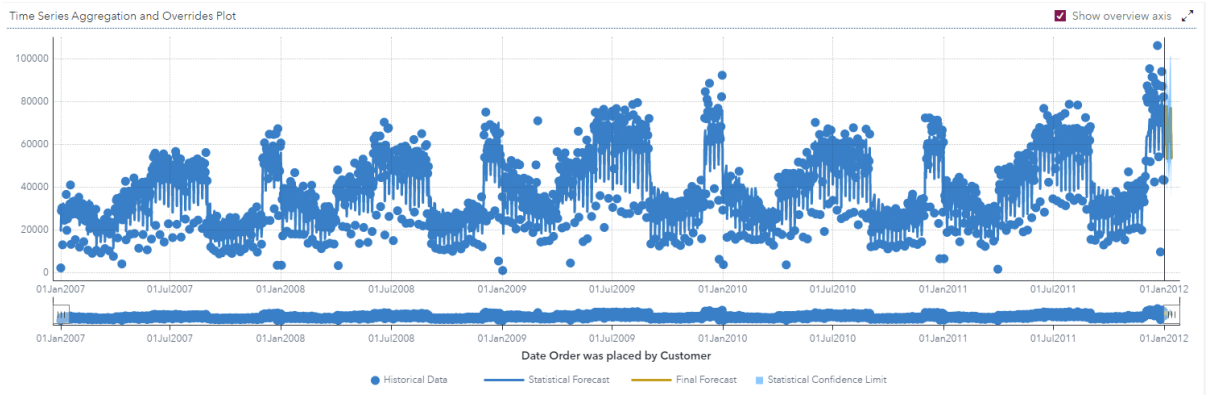

Figure 6 Time Series plot

Figure 6 shows us the time series plot of the orders. We can see there is a clear seasonal trend in the plot. The orders high during middle of every year in the months from May to August and then again at the end of the year in December. We see that the forecast also follows the same trend over the period of 12 days that have been forecasted for in figure 7.

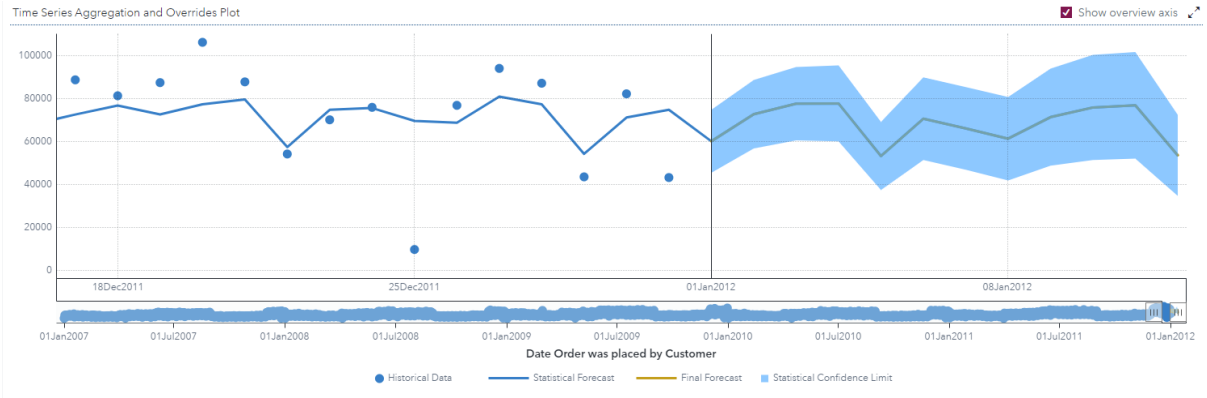

Figure 7 Forecast Model

The yellow line shows the forecasted model, and the shaded blue region shows the upper and lower limit of the forecast with 95% confidence.

Next, we are going to create a hypothetical scenario to override the current forecast. The scenario is if the profit increase by 10% for the day from January 1 to 6, 2012 and by 4% from January 7 to 12, 2012. We enter the same values in the override calculator. We get the follow result.

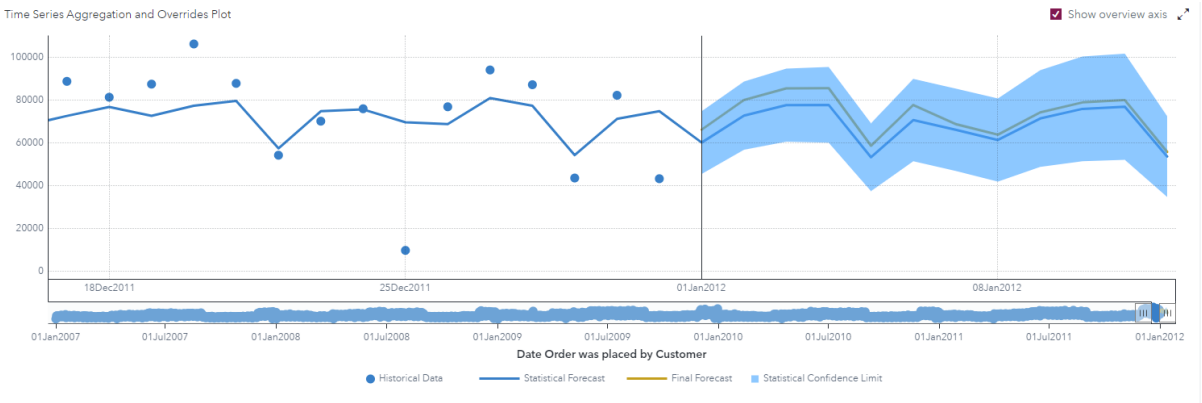

Figure 8 Forecast after override

In figure 8 we can see the blue line shows the original forecasting model and yellow line shows the forecast model after the override changes. As can be noticed, the difference between the yellow and blue line reduces from January 7, 2012 which is due to jump from 10% increase to 4% in profits.

The forecast values for 12 days are given in the table below.

Date	Statistical Forecast	Final Forecast	Lower Confidence Level	Upper Confidence Level
1st January 2012	60,141.61	66,155.77	45,547.43	74,735.79
2nd January 2012	72,711.92	79,983.11	56,825.85	88,597.98
3rd January 2012	77,611.64	85,372.80	60,571.00	94,652.28
4th January 2012	77,725.93	85,498.52	60,069.46	95,382.39
5th January 2012	53,318.63	58,650.49	37,547.71	69,089.54
6th January 2012	70,616.37	77,678.01	51,437.97	89,794.78
7th January 2012	66,109.64	68,754.02	46,926.44	85,292.84
8th January 2012	61,343.22	63,796.95	41,957.06	80,729.39
9th January 2012	71,354.80	74,208.99	48,851.18	93,858.41
10th January 2012	75,835.23	78,868.64	51,411.51	100,258.95
11th January 2012	76,835.68	79,909.11	52,104.38	101,566.99
12th January 2012	53,619.54	55,764.32	34,784.46	72,454.61

From the table above we get a clearer picture of how the forecasting model is working and what values have been predicted which is hard to read in the graph. We also see how the override function has worked and the new values.

Using this forecast model to forecast for longer time period will help retailers to order stock on time and keep products ready in time for the peak demand. This will not only help the supply chain running smoothly but also maximizing profits for everyone.