Part A

Hedonic House Price Model

Executive summary

A real estate group wanted to develop a hedonic model of housing prices by using the AMESHOUSING3 dataset to predict selling prices. To do this, the main goal is to identify the variables that affect the price of a house. First, through literature review, this report finds the most suitable variable selection method, model, and optimization method for this case. After determining the method, the existing 300 data are described, analyzed and processed for the subsequent application of the model. The variables with high correlation were eliminated by the correlation matrix to avoid the influence on the prediction results. The remaining data was then put into the multiple linear regression model to get the parameters of each variable and the predicted fit. The obtained model has a high R-square value, which can provide a good reference for real estate groups. The prediction model shows that foundation type, whether there is a garage and overall material and finish of the house are the three most important variables for the house price, and other variables also have a certain impact on the prediction.

Introduction

The existence of the real estate industry is the basic condition for improving people's living standards. The stable development of the real estate industry can not only bring significant profits to the industry, but also indirectly improve the local economic level and quality of life. Therefore, for real estate, it is very important to analyze the trend of house prices and the factors that lead to house price changes.

This report uses AMESHOUSING3, a data set from a leading real estate group. This dataset contains the characteristic data of 300 houses with more than 20 variables, which include the characteristic information and sales information of each house. To assist the real estate group in identifying the variables affecting housing prices in the hedonic housing price model, the AMESHOUSING3 dataset will be visualized using the SAS Viya tool, including correlation and regression analyses, in the following report. This paper will use multiple linear regression models to predict the price of houses. In the following report, the methodology for creating and selecting models will first be described, and then the models created will be evaluated and explained.

Literature review

This part will review the literature related to predicting housing price according to the model, variable selection, data processing and other methods that need to be used in this report and mention the same and different parts.

Amri & Tularam (2012) made a comparison between linear and nonlinear models in their research on predicting the housing price of Bathurst city. It is found that although the nonlinear model performs better, the linear model achieves the same exact effect by selecting the data in the correct way. The research of Amri & Tularam (2012) is similar to the research direction of this paper, but it has more complex data structure and larger data quantity. When it is mentioned in the study that more in-depth analysis is needed, the nonlinear method may be the better choice. However, since this study does not do too in-

depth analysis, the linear method may be more efficient.

Manasa, Gupta & Narahari (2020) used a variety of regression models, including multiple linear regression models, in their prediction of housing sales prices in cities such as Bangalore. The independent variables in the study of Manasa et al. are similar to this study, and the data processing methods are also similar. Just like the study in this report, there were a lot of missing data in their study, so they chose to eliminate such data. In addition, to fit the linear regression model, Manasa et al. transformed the variables with categorical features into numerical variables. This may be a useful reference for this report.

Ravikumar (2017) implemented hedonic regression model in the process of forecasting the real estate market and housing prices by using various algorithms. Ravikumar(2017) used the method of Limsombunchai (2004) to build the model in his study, that is, hedonic multiple regression model was selected for price prediction, and variables with multicollinearity were excluded from the model in model creation. This point of view also applies to this study because multicollinearity can affect the prediction results.

In the study of Alfiyatin, Febrita & Taufiq et al. (2017), the factors affecting housing price were considered as physical conditions, concepts, and locations. Based on these three factors, they used regression analysis and particle swarm optimization in the report to predict the price of houses. From the process of regression analysis, it can be found that Alfiyatin et al. (2017) did not convert variables into categorical variables when faced with variables such as whether there is public transport, but directly used 0 and 1 to represent them.

Methodology

Data processing

First, in the descriptive analysis section, all variables will be included and classified. These variables will be classified as continuous variables and categorical variables.

The variables will then be pre-processed. In this step, the skewness of the

continuous variable is evaluated to see if they need to be logarithmic or deleted directly. Since taking the logarithm does not change the correlation between the data, the purpose is only to reduce the value larger than the median by a certain proportion to form a normally distributed data.

Model selection

This report plans to create regression models to make predictions. Since there is no dichotomous dependent variable in AMESHOUSING3 data set, logistic regression model is not considered in the analysis process, and only multiple linear regression models will be created and evaluated.

Variable selection

In the process of creating multiple linear regression models, the dependent variables will be selected as continuous variables. In addition, in the selection of independent variables, on the one hand, all continuous variables will be used to create a correlation matrix and correlation analysis. If two continuous variables are highly correlated, that is, when the correlation is greater than 0.7, there may be multicollinearity between the variables, which will lead to unstable prediction. Thus, only one of the continuous variables will be taken.

On the other hand, since categorical variables cannot be analyzed by correlation matrix, this report will judge the correlation between two variables and the existence of potential multicollinearity through logic.

Model evaluation

In the process of evaluating the established linear regression models, this report will use the assessment plot to judge the results and errors of the prediction. In this report, Backward will be used to perform multiple calculations and comparative tests on the established regression model until all the remaining independent variables in the model are statistically significant. R-square will be used to judge the degree of fit. The closer the value is to 1, the better the degree of fit, and the residual plot is used to visually show the fit of independent variables and dependent variables. P value is an index used to test the reliability of the model. This report will use the P value 0.01 as the standard to judge whether the results of the regression model are significant.

Analysis

Descriptive analysis

By using the SAS Viya tool, continuous variables are found in the measurement details, and all data except the score data are not missing. Since all 300 data of score are missing, this variable will not be referenced in the subsequent regression analysis. See *Appendix 1* for a table of continuous variables. Since PID will not be included in the following regression models, descriptive analysis of categorical variables will exclude it. The categorical variables in the AMESHOUSING3 dataset included foundation type, garage type, heating quality, central air conditioning, masonry veneer, lot shape and style of dwelling.

Data processing

The partial value of each continuous variable can be viewed according to the continuous variable table obtained in the descriptive analysis. This report was intended to log all variables with large partial values, but since virtually all continuous variables lie between -1 and 2, additional logarithmic processing is no longer necessary. A group variable transformation will then be performed on some continuous variables. Due to the nonlinear relationship between the year/month/season and the house sale price, we choose to classify and convert these three items here. The transformed categorical variables are shown in *Appendix 2*.

Correlation analytics

By establishing a correlation matrix containing all continuous independent variables, we can clearly see the correlation between variables. As shown in *Appendix 3*, to avoid the influence of multicollinearity on the prediction model, we eliminate one of the two variables with strong relationship (that is, correlation greater than 0.7) in this matrix to obtain a new correlation matrix. The correlation between variables in the new correlation matrix is less than 0.7. For the 8 categorical variables contained in the current dataset and the 3 categorical variables obtained by transformation, the relationship between each variable is not so high logically, but the Style of dwelling has a unique duplication, so we only chose one of them in the regression model.

Regression analysis

In the linear regression model, Sale price in dollars and Natural log of the Sale price are respectively selected to establish two models for comparison after the dependent variable is completed, as shown in *Appendix 4*, from which we intend to select the optimal model. By using Backward for these two linear regression models, the independent variables with P-value greater than 0.01 and the least significant in the model are successively eliminated to obtain the completed model in *Appendix 5* and *6*. By observing the fitting degree of each value and different variables in the two linear regression models, it was found that the linear regression model with Natural log of the sale price as the dependent variable was more in line with the requirements of this analysis report, so it was selected as the final analysis model.

In this model, an R-square value of 0.8967 indicates that 89.67% of the prediction results can be explained by the model. The closer this value is to 1, the more explanatory power the model has. It can be seen from the fit summary in *Appendix 6* that 297 out of the 300 total data are used in this model. The parameters of each variable used in the final model are shown in the Parameter Estimates table in *Appendix 6*. Through the parameters in the table, detailed explanations of each variable are given as follows:

- **Foundation Type**: Based on Foundation Type Concrete/Slab, when Foundation Type is Cinder Block, the house price increases by 0.70% compared with Concrete/Slab. When the Foundation Type is Brick/Tile/Stone, it has the biggest impact on the house price, which is 7.57% more than Concrete/Slab.
- **Garage attached or detached**: Based on Garage attached or detached NA, when garage is attached, a home sell for 11.50% more. Detached When Garage is a detached class, home prices are up 11.46%.
- **Above grade (ground) living area square feet**: With other variables unchanged, a 1% increase in Above Grade (ground) Living area square feet will increase the selling price by 0.29%.
- **Age of house when sold, in years**: A 1% increase in Age of House when sold reduces the sale price by 0.32%, holding other variables constant.

- **Basement area in square feet**: A 1% increase in Basement Area in Square Feet increases the sale price by 0.016%, holding other variables constant.
- **Lot size in square feet**: With other variables unchanged, a 1% increase in Basement Area in Square feet will increase the house price by 0.00012%.
- **Number of fireplaces**: A 1% increase in the number of Fireplaces increases the sale price of the home by 5.83%, holding other variables constant.
- **Number of full bathrooms**: A house sale price will increase 3.95% from the bathrooms if the number of full bathrooms increases 1%, assuming all other variables are constant.
- **Overall condition of the house**: A 1% increase in Overall Condition of the House, holding other variables constant, would increase the selling price of the house by 5.79%.
- **Overall material and finish of the house**: Holding other variables constant, a 1% increase in Overall Material and Finish of the house would increase the selling price of the house by 7.42%.

By observing the residual plot in *Appendix 6*, the residual distribution of this linear regression model is considered, and it is found that there is no obvious trend in the residual distribution. On the other hand, the predicted average curve shown in the assessment plot of the regression model almost coincides with the observed average curve. Although some curves show some errors, they are all within the acceptable range, indicating that the model has a good fitting degree and can be used to predict the house selling price.

Conclusions

The above report uses multiple linear regression models to predict home selling prices in the AMESHOUSING3 dataset. By comparing the two established models, the model with a better dependent variable is chosen to make the prediction more accurate.

Through observation and analysis of the finally selected linear regression model, it is found that foundation type is an important predictor as a categorical variable.

When the foundation type is brick/tile/stone, the selling price of the house increases significantly. Both types of garages have a positive impact on house prices, and the impact is similar.

In predicting the price of a house, real estate groups need to focus on the number of fireplaces, number of full bathrooms, overall condition of the house, overall material and finish of The four continuous variables, the house, have a great positive impact on the selling price of the house. Above grade (ground) living area square feet, Age of house when sold, Basement area in square feet, Lot size in square feet has no greater impact on house selling price than the first four and can be used as a secondary observation factor.

References

Alfiyatin, A.N., Febrita, R.E., Taufiq, H. & Mahmudy, W.F., 2017. Modeling house price prediction using regression analysis and particle swarm optimization case study: Malang, East Java, Indonesia. International Journal of Advanced Computer Science and Applications, 8(10).

Amri, S. & Tularam, G.A., 2012. Performance of multiple linear regression and nonlinear neural networks and fuzzy logic techniques in modelling house prices. Journal of Mathematics and Statistics, 8(4), pp.419-434.

Limsombunchai, V. (2004). House price prediction: hedonic price model vs. artificial neural network, New Zealand Agricultural and Resource Economics Society Conference, pp. 25–26.

Manasa, J., Gupta, R. & Narahari, N.S., 2020, March. Machine learning based predicting house prices using regression techniques. In 2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA) (pp. 624-630). IEEE.

Ravikumar, A.S., 2017. Real estate price prediction using machine learning (Doctoral dissertation, Dublin, National College of Ireland).

Part B

Profit Forecasts

Executive summary

This report uses SAS Viya forecasting tools to forecast the profit in the PRODUCTANALYSIS dataset. It will start by assigning the appropriate roles to the required variables. The reasons and methods for choosing variable roles will be explained. After selecting appropriate variables, the model is selected. In this study, three models are selected, namely Naive Model, Auto-forecasting Model and Hierarchical Forecasting Model. They will be placed in the same pipeline and compared. The comparison process is automatically completed by SAS Viya. This report will observe and analyze the comparison results and select the optimal forecasting model, that is, the hierarchical forecasting model with relatively small WMAE. In addition, this presentation presents the predicted results and simulates a special case for the calculation of override.

Methodology

This section will discuss the methods used to select variables and models during the research process and explain how to compare models.

Variable selection

With Order_Date selected as the Time role and Profit selected as the dependent variable, roles are assigned to the appropriate variables. First, the explanatory variables for profit should be selected. In the PRODUCTANALYSIS data set, this report selects variables that can directly affect Profit, that is, can obtain Profit value through calculation. They are Cost, Discount, Quantity and RetailPrice, respectively. As for the selection of the BY variable, the geographical location is not a good choice, so the product attributes are selected as the BY variable in this report, including Product_Line, Product_Category, and Product_Group. In addition, when selecting the Reconciliation Level, we choose Product_Group by using bottom-up approach in order to come out with more accurate forecasting. For Naïve Model, the type is changed to moving average with the window size setting to 7, so that the forecast can be the average of each week. The variables are detailed in *Appendix 7*.

Model Comparison

In order to find a more suitable forecasting model, this study chooses Auto-forecasting, Naive and Hierarchical Forecasting Model for comparison. In the model building process, these three forecasting models will be put in the same pipeline, so that the comparison results can be more clearly displayed. Moreover, they use the same variables to make sure the comparison is accurate. The process of pipeline running is automatically done by SAS, and the results will be displayed in Model Comparison, which directly shows the values of WMAE and WMAPE and the champion Model considered by SAS. Based on this result, this study will select the final model for use.

Results

Model selection

Using the approach described in the previous section, a pipeline containing both auto-forecasting, Naive and Hierarchical forecasting models as outlined in *Appendix 8* has been created. After running the pipeline, we get the Model Comparison in *Appendix 9*.

Each model is described as follows:

- **Auto-forecasting Model:** It shows that the WMAE value is 412.6171 and the WMAPE value is 52.0858. Moreover, it was selected as the champion model by SAS.
- **Naive Model:** It is shown to have a WMAE value of 394.3075 and a WMAPE value of 56.2174.
- **Hierarchical Forecasting Model:** It is displayed that the WMAE value is 560.2411 and the WMAPE value is 87.2368.

Because SAS chose the Auto-forecasting Model as the champion, thus it is our preferred Model.

Forecast and Override

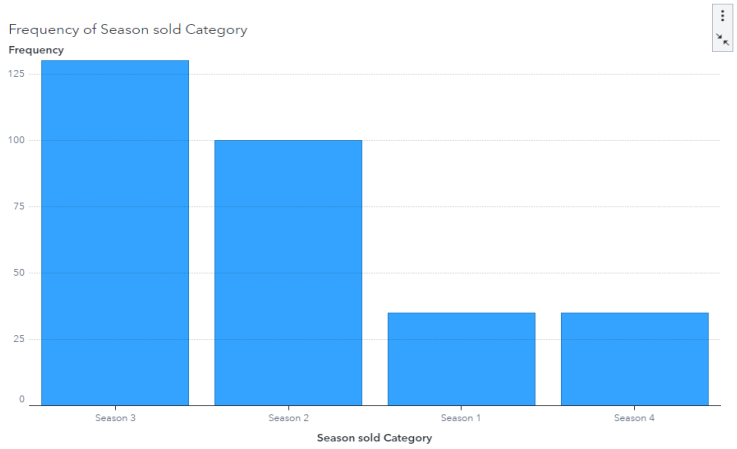
The prediction results of the model in this study are presented in *Appendix 10*. To test the calculation of Override, we drew up a special case for the prediction process. Suppose that starting in January 2012, because of a sudden increase in the demand for the product, the company decides to take the opportunity to adjust the price. This adjustment will result in a 3% increase in the final profit forecast. The final prediction results after calculation were obtained by the Override calculator of SAS, as shown in *Appendix 10*.

Appendix

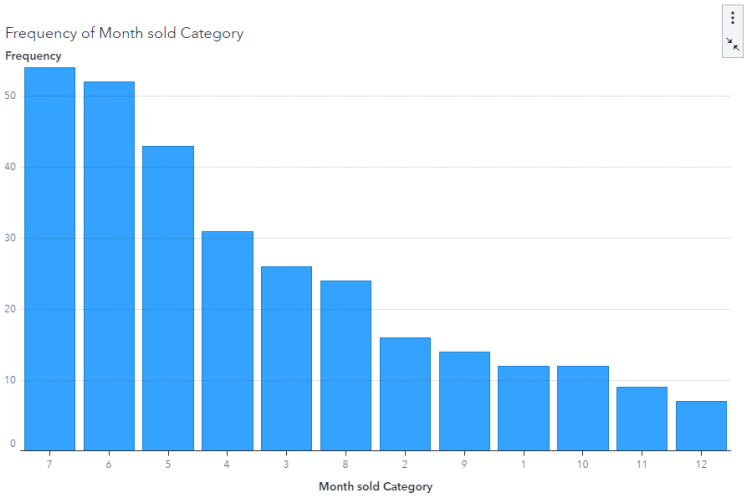
Appendix 1: Continuous Variables

Variable	Skewness	Missing Value
Above grade (ground) living area square feet	-0.3905	0
Age of house when sold, in years	0.1953	0
Basement area in square feet	-0.5477	0
Bedrooms above grade	-0.7203	0
Lot size in square feet	1.0093	0
Month sold (MM)	0.2196	0
Natural log of the sale price	-0.9443	0
Number of fireplaces	1.1144	0
Number of full bathrooms	0.5397	0
Number of half bathroom	1.3819	0
Original construction year	-0.1821	0
Overall condition of the house	0.4044	0
Overall condition of the house	-0.5838	0
Overall material and finish of the house	-0.3144	0
Overall material and finish of the house	-0.5748	0
Sale price > $175,000	1.9703	0
Sale price	0.2973	0
Score	-	300
Season when house sold	-0.1578	0
Size of garage in square feet	-0.3738	0
Total area of decks and porches in square feet	1.4534	0
Total number of bathrooms (half bathrooms counted 10%)	0.5420	0
Year sold (YYYY)	0.0521	0

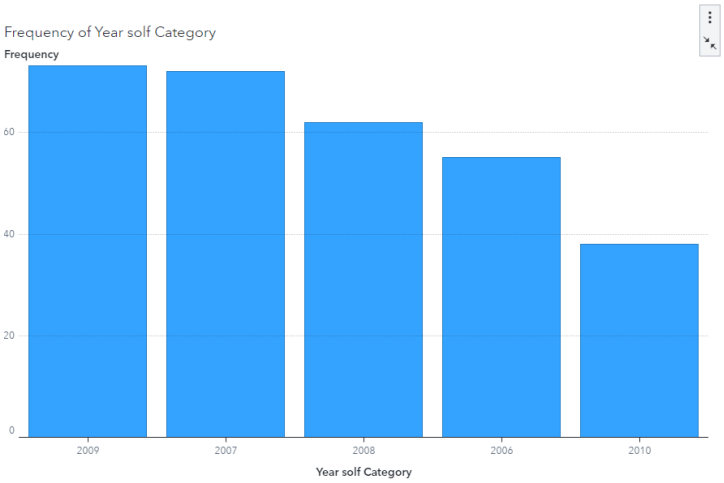
Appendix 2: Grouping Variables



Season sold category

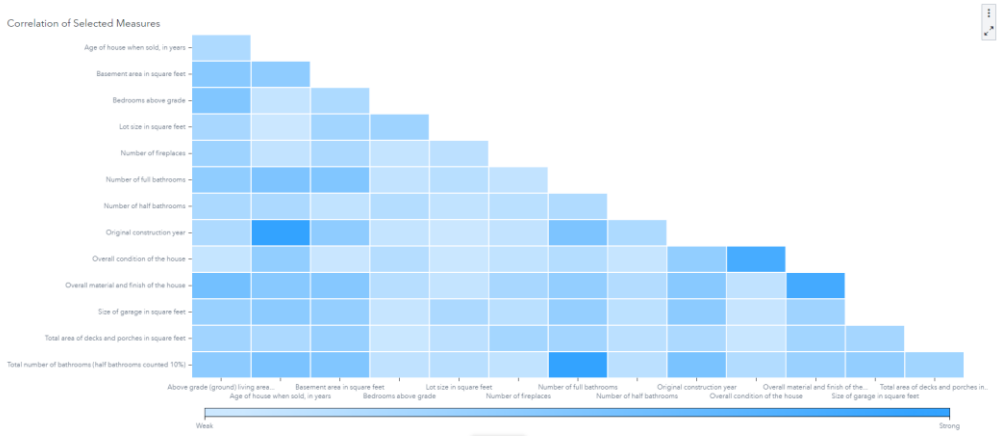


Month sold category

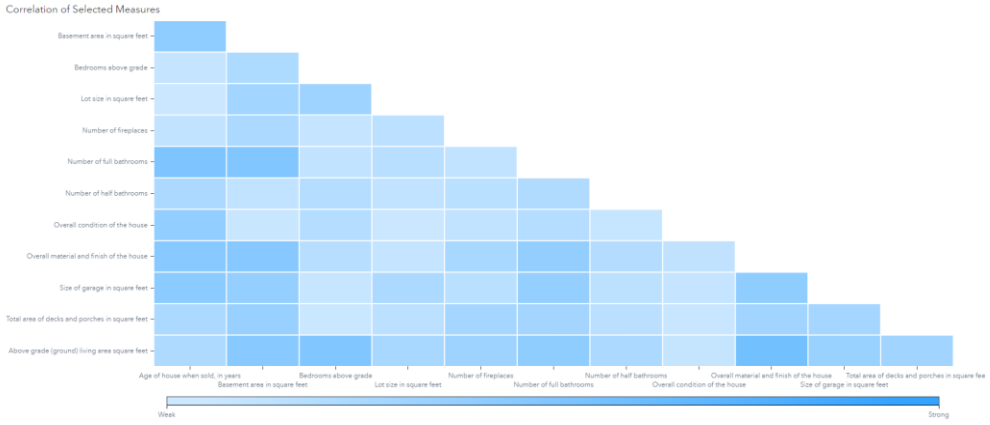


Year sold category

Appendix 3: Correlation Matrix

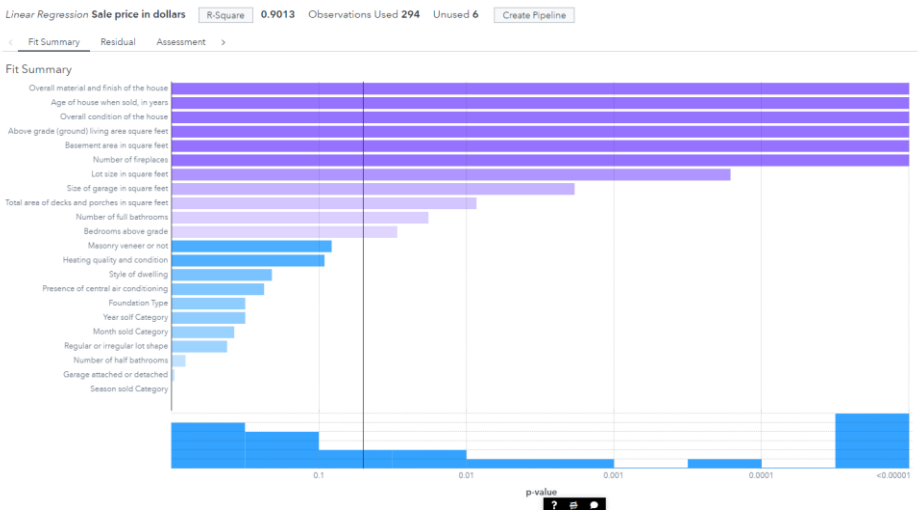


Correlation between all variables



Correlation between variables except strong relationship

Appendix 4: Linear regression model before Backward

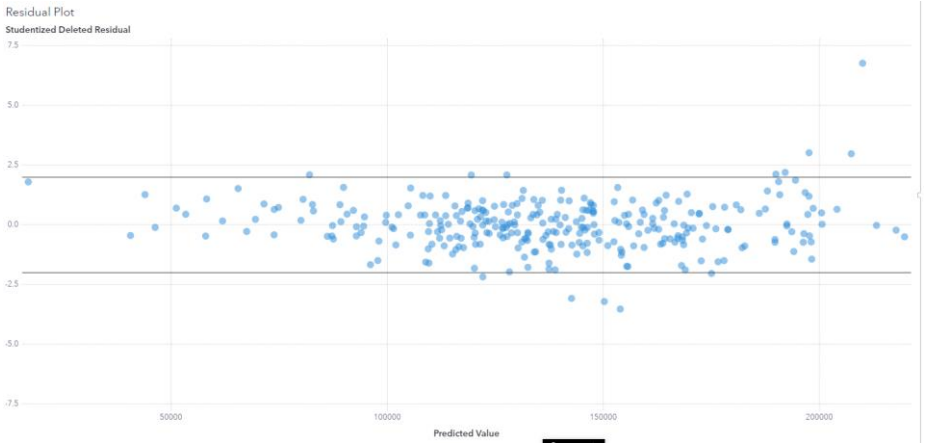


Linear regression model (Sale price in dollars)

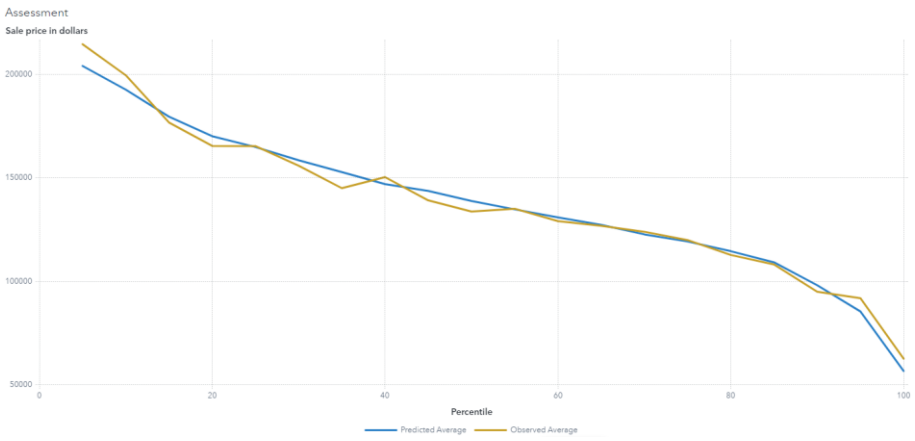
Linear regression model (Natural log of the sale price)

Appendix 5: Linear regression model after Backward (Sale price in dollars)

Fit summary



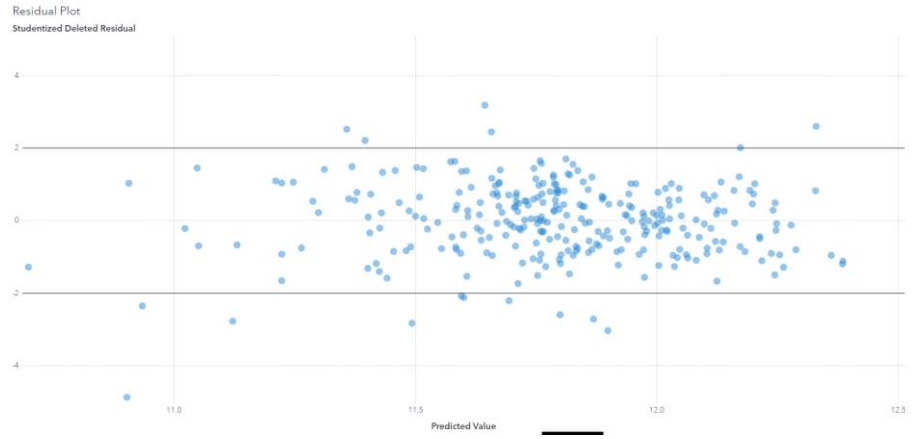
Residual plot



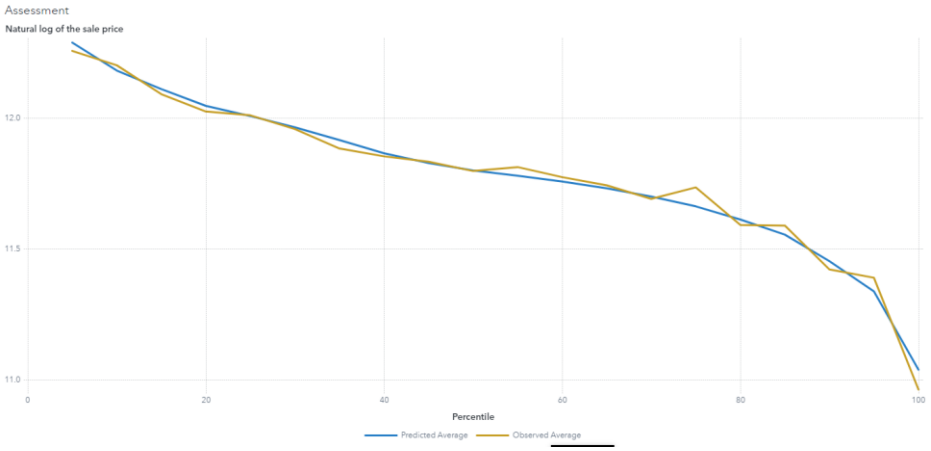
Assessment plot

Appendix 6: Linear regression model after Backward (Natural log of the sale price)

Fit summary



Residual plot



Assessment plot

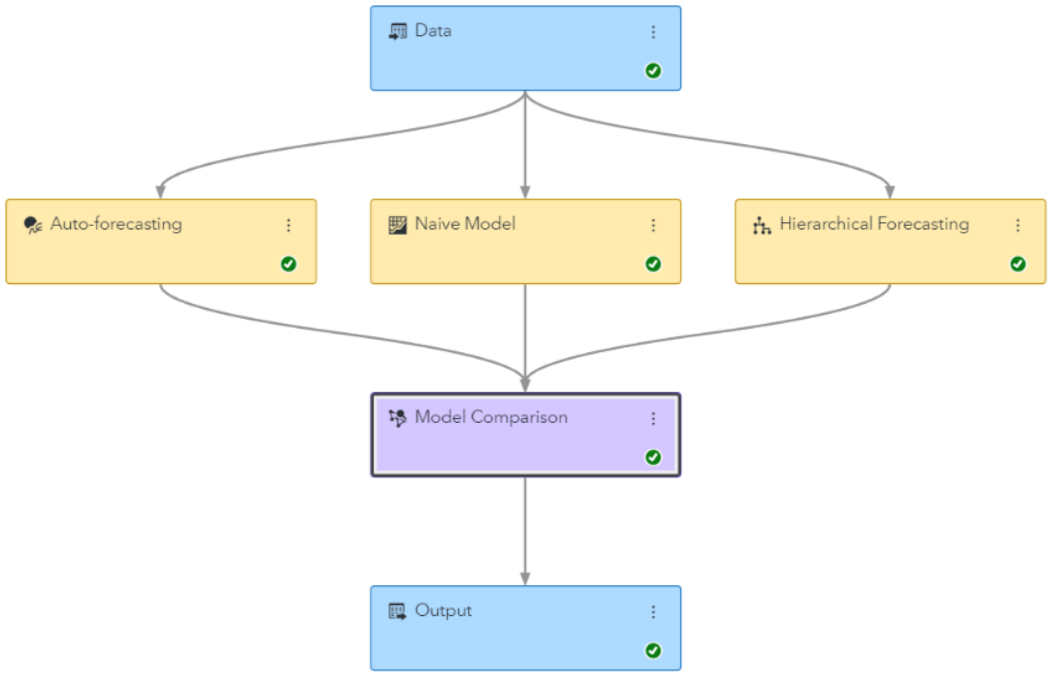
Parameter	Estimate	t Value	Pr > |t|
Intercept	10.45241488	216.652275	0.00000
Foundation Type Brick/Tile/Stone	-0.075662592	-3.043242207	0.00256
Foundation Type Cinder Block	-0.007421466	-0.503810073	0.61479
Foundation Type Concrete/Slab	0		
Garage attached or detached Attached	0.115004234	5.114480309	0.00000
Garage attached or detached Detached	0.114675194	5.468024441	0.00000
Garage attached or detached NA	0		
Above grade (ground) living area square feet	0.000286578	8.34892741	0.00000
Age of house when sold, in years	-0.003221351	-8.81553405	0.00000
Basement area in square feet	0.00016366	7.710446031	0.00000
Lot size in square feet	1.17845E-05	6.155070135	0.00000
Number of fireplaces	0.058275375	5.264229759	0.00000
Number of full bathrooms	0.039506243	3.519032442	0.00050
Overall condition of the house	0.057934741	10.29699997	0.00000
Overall material and finish of the house	0.07417946	10.06610363	0.00000

Parameter Estimates

Appendix 7: Forecasting Variables Roles

Variable Name	Type	Role	Hierarchy Aggregation
Order_Date	Numeric	Time	
Profit	Numeric	Dependent	Sum of values
Product_Category	Character	BY Var	
Product_Group	Character	BY Var	
Product_Line	Character	BY Var	
Cost	Numeric	Independent	Average of values
Discount	Numeric	Independent	Average of values
Quantity	Numeric	Independent	Average of values
RetailPrice	Numeric	Independent	Average of values

Appendix 8: Pipelines



Parameter Estimates

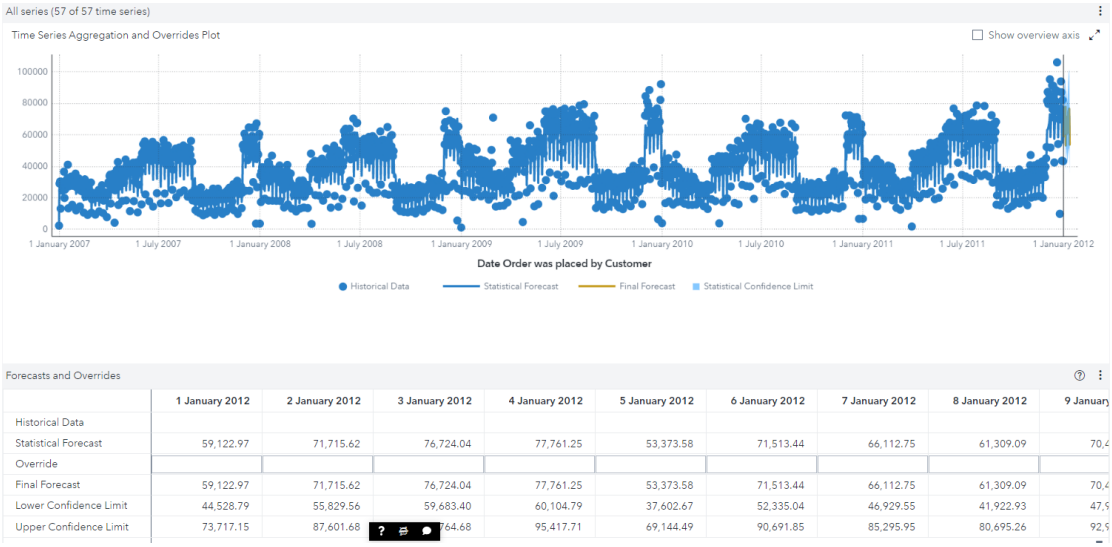
Appendix 9: Model Comparison

Model Comparison

Champion	Model Name	Status	WMAE	WMAPE
★	Auto-forecasting	Successful	412.6171	52.0858
	Hierarchical Forecasting	Successful	394.3075	56.2174
	Naive Model	Successful	560.2411	87.2368

Reconciliation Level: Product_Group

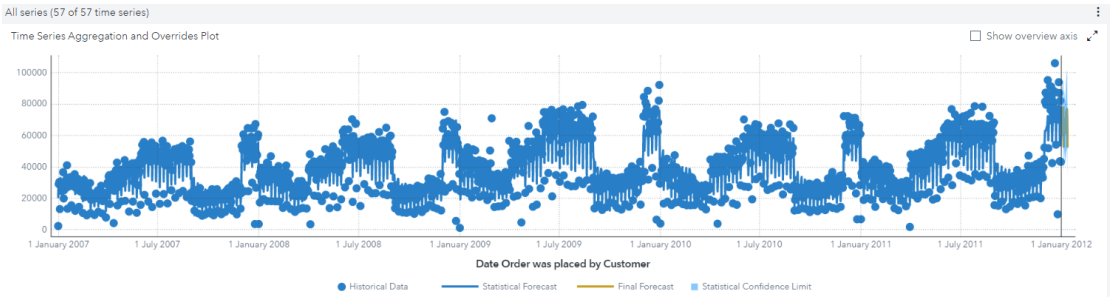
Appendix 10: Forecast and Override



Forecasts and Overrides

	1 January 2012	2 January 2012	3 January 2012	4 January 2012	5 January 2012	6 January 2012	7 January 2012	8 January 2012	9 January
Historical Data									
Statistical Forecast	59,122.97	71,715.62	76,724.04	77,761.25	53,373.58	71,513.44	66,112.75	61,309.09	70,4
Override									
Final Forecast	59,122.97	71,715.62	76,724.04	77,761.25	53,373.58	71,513.44	66,112.75	61,309.09	70,4
Lower Confidence Limit	44,528.79	55,829.56	59,683.40	60,104.79	37,602.67	52,335.04	46,929.55	41,922.93	47,9
Upper Confidence Limit	73,717.15	87,601.68	764.68	95,417.71	69,144.49	90,691.85	85,295.95	80,695.26	92,9

Before Override calculation



Forecasts and Overrides

	1 January 2012	2 January 2012	3 January 2012	4 January 2012	5 January 2012	6 January 2012	7 January 2012	8 January 2012	9 January
Historical Data									
Statistical Forecast	60,844.49	72,170.67	76,840.22	77,893.38	52,957.71	70,472.81	66,572.79	61,639.77	70,4
Override									
Final Forecast	60,844.49	72,170.67	76,840.22	77,893.38	52,957.71	70,472.81	66,572.79	61,639.77	70,4
Lower Confidence Limit	46,250.31	56,284.61	59,799.57	60,236.92	37,186.80	51,294.40	47,389.59	42,253.61	47,9
Upper Confidence Limit	75,438.67	88,056.73	880.86	95,549.84	68,728.63	89,651.21	85,755.99	81,025.93	93,0

After adjustment +3% Override calculation