Predictive Analytics with California Airbnb Dataset/ SAS Enterprise Miner

Didem B. Aykurt

Colorado State University Global

MIS530; Predictive Analytics

Dr.Jennifer Catalano

March 12, 2023

Table of Contents

Predictive Analysis of California Airbnb Dataset

Introduction

A privately owned multinational corporation headquartered in San Francisco, Airbnb, Inc. runs an online trade and accommodation business that may be accessed through its apps and web apps. Subscribers of the website can use it to book or provide accommodation, generally guesthouses or travel opportunities.

Brian Chesky and Joe Gebbia built a company for a cash grab. First, they bought mattresses and rented part of their San Francisco apartment. Next, they had the first guest attending a design conference; after a successful weekend, they added a third founder, Nate Blecharczyk, in 2008. Twelve years later, Airbnb is the world's most significant tourism to traveling days and has over 5.6 million active listings in 220 countries, at least 100,000 cities, 4M hosts, and over 1B guests. Airbnb had a value of $86.5B at IPO. They were now selling at $146 per share on opening day.

My family and I are the best Airbnb customers when we vacation. We always use the Airbnb web to choose the best price at a beautiful place. However, that company name is so unique that three words make the company name Air, Bed, and Breakfast. That is why I chose the company and want to see how the company increases loyalty. Airbnb hosts’ pricing strategy is the key to having long-run resemble, such as charging a lower price to entice more customers and executing higher residence rates instead of a short-run approach. Airbnb listing price is a significant risk faced when entering the market. As a result, businesses must study the elements contributing to listing prices and understand the perfect daily price to charge and its effects.

To examine the relationship between room type and price, I will use SAS Enterprise Miner to analyze price and forecast the key elements contributing to a higher occupancy rate. I then compared those findings to the pricing amount. Additionally, I conducted a descriptive study to examine a few crucial factors that would be very beneficial for business, such as:

1. What California neighborhoods are the most popular for Airbnb rentals?
2. What is the relation between Airbnb guests' most local neighborhood area, room type, property, and price?

For the analysis, a public database from the Airbnb platform was utilized. The dataset offers details on the characteristics of homes, review ratings, comments, and the availability of more than 10,000 listings in 2019. The Airbnb data was employed to execute visualizations, and SAS additionally carried out linear regression to identify the elements influencing higher ratings. SAS was also used to analyze consumer reviews.

I have chosen to explore the Bay Area, CA-Airbnb Data (UPDATE 2020) CSV dataset, an open data source available on Kaggle and has been updated June 12th, 2020, and the variables required to address the business problem.

The research aims to build price recommendations, factors affecting residence rates, and the hypothesis that room type plays a significant role in booking the Airbnb analysis in California. This report is divided into three milestones. Milestone 1 introduces and defines the business problem of Airbnb, the dataset, etc. Milestone Project two descriptive statistics describe four minimum business problems and create alternate and null hypotheses for each business question. It also includes testing the ideas with an appropriate statistical test. Finally, milestone project 3 performs a predictive analysis technique, compares the different models' performance, and identifies the best model for the CA-Arbnb.csv dataset.

The business problem of Airbnb states that we can say which neighborhood has the highest price range for the listings. From this, we can find out that the solution to the problem is to regulate the price of areas or room types.

CA-Airbnb.csv data set has 7221 observations of 106 variables. Therefore, I will use 20 variables or columns, including 15 numerical and five-character variables.

Data Description of Listings, Calendar, and Reviews

Variable	**Description**
ID	Listing id of the property
Name	Name of the property
Host_Id	Id of the property host
host_name	Name of the host property
neighborhood_cleansed	The neighborhood of the property
Latitude	Location of the Latitude
Longitude	Location of the Longitude
Property_type	Type of the property

room_type	Type of the room
bathrooms	Total number of bathrooms
bedrooms	Total number of bedrooms
beds	Total number of beds
price	Price of the property
Min_nights	Minimum number of nights required to book
number_of_reviews	Total number of reviews
availability_365	Availability of property
last_review	Date of the last review
review_scores_rating	The total score of the review rating
Calculated_host_ _Listings_count	Total listings the host has
Reviews_per_month	Average Number of reviews in a month

Import CA Airbnb Dataset into SAS Enterprise Miner

Figure 1

Import CA-Airbnb Dataset into SAS Enterprise Miner.

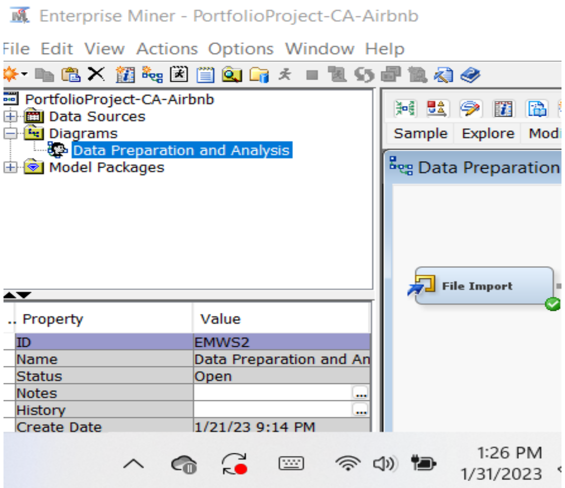


Figure 2

Output window of File Import node for CA-Airbnb Dataset.

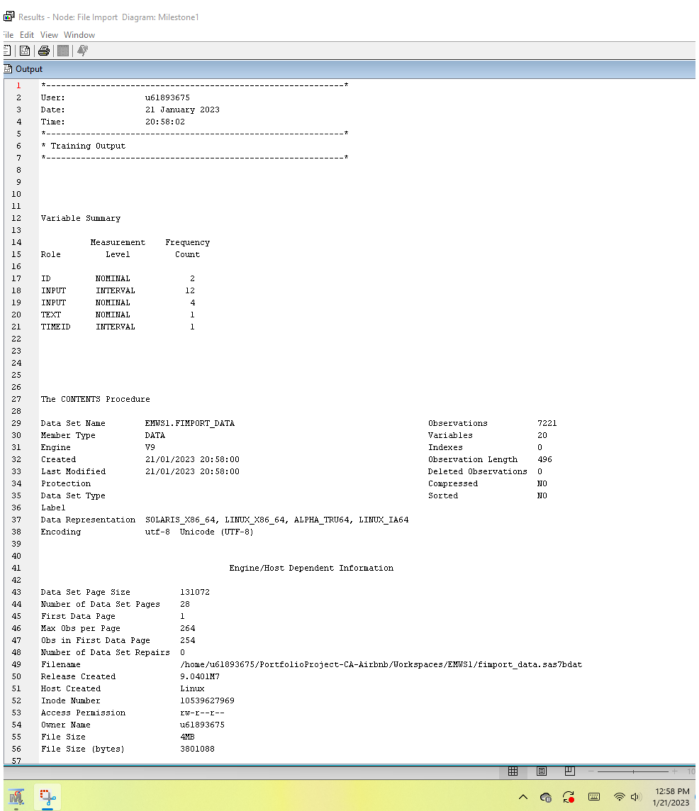


Figure 3

Output of Variables Alfabetics List.

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
16	availability_365	Num	8	BEST.		availability_365
10	bathrooms	Num	8	BEST.		bathrooms
11	bedrooms	Num	8	BEST.		bedrooms
12	beds	Num	8	BEST.		beds
19	calculated_host_listings_count	Num	8	BEST.		calculated_host_listings_count
3	host_id	Num	8	BEST.		host_id
4	host_name	Char	35	$35.	$35.	host_name
1	id	Num	8	BEST.		id
17	last_review	Num	8	MMDDYY10.		last_review
6	latitude	Num	8	BEST.		latitude
7	longitude	Num	8	BEST.		longitude
14	minimum_nights	Num	8	BEST.		minimum_nights
2	name	Char	281	$281.	$281.	name
5	neighbourhood_cleansed	Char	20	$20.	$20.	neighbourhood_cleansed
15	number_of_reviews	Num	8	BEST.		number_of_reviews
13	price	Num	8	NLMNY15.2		price
8	property_type	Char	18	$18.	$18.	property_type
18	review_scores_rating	Num	8	BEST.		review_scores_rating
20	reviews_per_month	Num	8	BEST.		reviews_per_month
9	room_type	Char	15	$15.	$15.	room_type

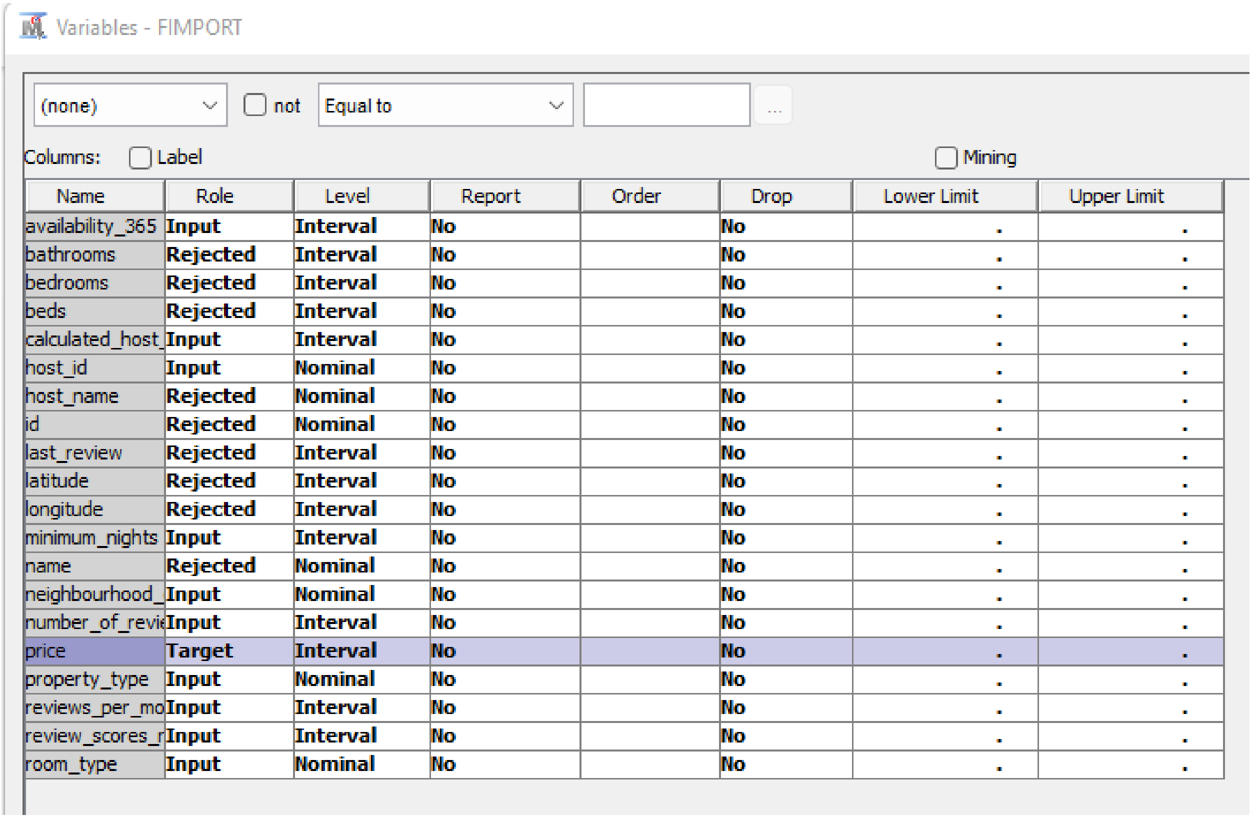
1:04 PM
3/4/2023

Editing Variables

Rejected a few variables: bathrooms,’ ‘bedrooms,’ ‘beds,’ 'id,’ 'host_name,’ and 'last_review' are unnecessary to address the business problem because these drop variables are irrelevant and insignificant to our investigation. So, instead, I set the target variable for the price on the Variables-FIMPORT window. At the end of the variable eliminated process, the CA-Airbnb dataset contains 7221 records with 11 different attributes including but not limited to availability_364, calculate_host_listing_count, host_id, minimum_nights, neigborhood_cleansed, room_type, and price.

Figure 4

Variables-FIMPORT- Update variables' Roles.



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
availability_365	Input	Interval	No		No	.	.
bathrooms	Rejected	Interval	No		No	.	.
bedrooms	Rejected	Interval	No		No	.	.
beds	Rejected	Interval	No		No	.	.
calculated_host_	Input	Interval	No		No	.	.
host_id	Input	Nominal	No		No	.	.
host_name	Rejected	Nominal	No		No	.	.
id	Rejected	Nominal	No		No	.	.
last_review	Rejected	Interval	No		No	.	.
latitude	Rejected	Interval	No		No	.	.
longitude	Rejected	Interval	No		No	.	.
minimum_nights	Input	Interval	No		No	.	.
name	Rejected	Nominal	No		No	.	.
neighbourhood_	Input	Nominal	No		No	.	.
number_of_revi	Input	Interval	No		No	.	.
price	Target	Interval	No		No	.	.
property_type	Input	Nominal	No		No	.	.
reviews_per_mo	Input	Interval	No		No	.	.
review_scores_r	Input	Interval	No		No	.	.
room_type	Input	Nominal	No		No	.	.

Data Partition

Created a training of 55% of the dataset to train or develop the model, and validation of 45% of the dataset will be used to validate it. To connect the Data Partition node to the File Import node.

Figure 5

Data Set Allocation.

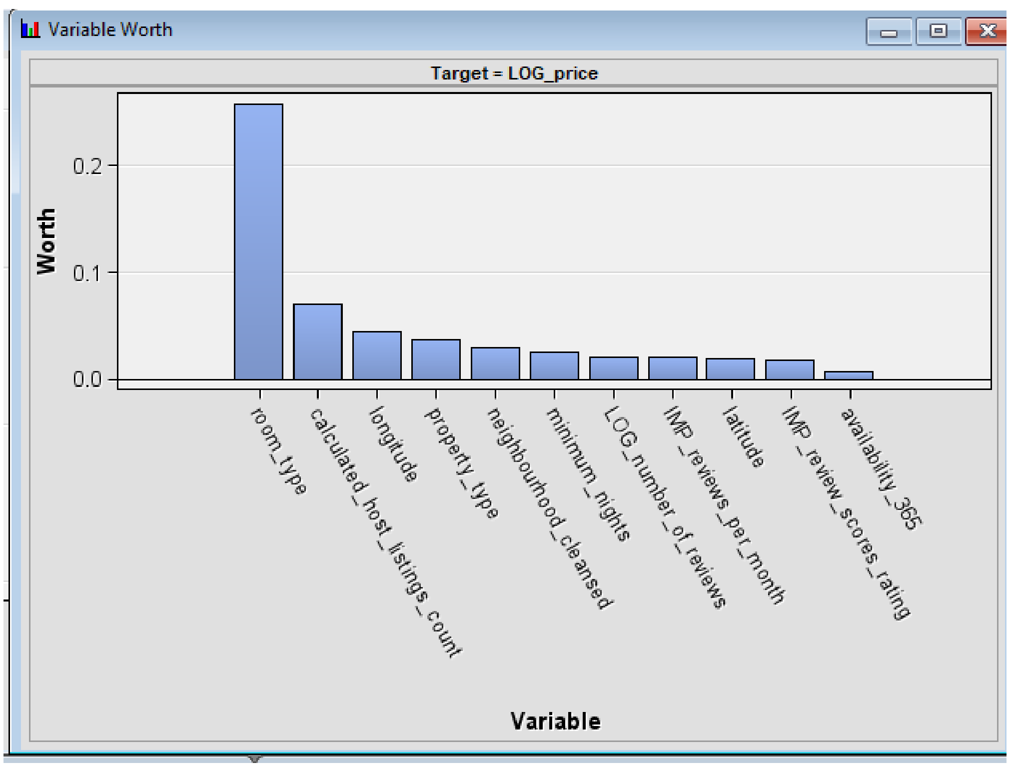
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	55.0
Validation	45.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	1/24/23 8:15 PM
Run ID	856b4e37-70ce-ad49-a

Data Exploration

The most worthwhile variables are room type, property type, and neighborhood. I will build hypothesis testing for room type and neighborhood variables to get more details about the price average.

Figure 6

The bar chart shows the Variable Worth window by price.



The highest percentage of room type was a private room at 47.83%, then an Entire home/apt at 47.51%.

Figure 7

Result of summary statistics Room type class variable frequency and percentage.

Distribution of Class Target and Segment Variables
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Level	Frequency Count	Percent
TRAIN	room_type	TARGET	Private room	3454	47.8327
TRAIN	room_type	TARGET	Entire home/apt	3431	47.5142
TRAIN	room_type	TARGET	Shared room	334	4.6254
TRAIN	room_type	TARGET	Hotel room	2	0.0277

The result of StatExplore shows that the highest percentage of room types by neighborhood is Hotel rooms in Palo Alto 50% and Santa Clara 50%.

Figure 8

Result of summary statistics Room type by neighborhood.

Class Variable Summary Statistics by Class Target
(maximum 500 observations printed)

Data Role=TRAIN Variable Name=neighbourhood_cleansed

Target	Target Level	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
OVERALL		16	0	San Jose	40.39	Palo Alto	11.67
room_type	Entire home/apt	16	0	San Jose	37.62	Palo Alto	15.29
room_type	Hotel room	2	0	Palo Alto	50.00	Santa Clara	50.00
room_type	Private room	15	0	San Jose	43.80	Sunnyvale	10.92
room_type	Shared room	11	0	San Jose	30.60	Santa Clara	20.77

The private room median at $67 is a private room, and the highest median at $170 is a hotel room. The cheapest room type is a shared room.

Figure 9

Result of summary statistics Room type by price.

Data Role=TRAIN Variable=price

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation
OVERALL		100	0	7221	10	10000	161.2871	352.069
room_type	Entire home/apt	168	0	3431	10	5500	232.9948	304.9958
room_type	Hotel room	170	0	2	170	199	184.5	20.5061
room_type	Private room	67	0	3454	10	10000	100.3891	392.0329
room_type	Shared room	30	0	334	15	2900	54.2994	180.9512

The highest percentage of the neighborhood is San Jose at 41.29%, then Palo Alto at 11%.

Figure 10

Result of summary statistics for Neighborhood frequency and percentage list.

Distribution of Class Target and Segment Variables
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Level	Frequency Count	Percent
TRAIN	neighbourhood_cleansed	TARGET	San Jose	1529	41.2908
TRAIN	neighbourhood_cleansed	TARGET	Palo Alto	423	11.4232
TRAIN	neighbourhood_cleansed	TARGET	Sunnyvale	376	10.1539
TRAIN	neighbourhood_cleansed	TARGET	Santa Clara	340	9.1817
TRAIN	neighbourhood_cleansed	TARGET	Mountain View	326	8.8037
TRAIN	neighbourhood_cleansed	TARGET	Milpitas	155	4.1858
TRAIN	neighbourhood_cleansed	TARGET	Unincorporated Areas	155	4.1858
TRAIN	neighbourhood_cleansed	TARGET	Cupertino	154	4.1588
TRAIN	neighbourhood_cleansed	TARGET	Campbell	63	1.7013
TRAIN	neighbourhood_cleansed	TARGET	Los Gatos	44	1.1882
TRAIN	neighbourhood_cleansed	TARGET	Los Altos	39	1.0532
TRAIN	neighbourhood_cleansed	TARGET	Saratoga	32	0.8642
TRAIN	neighbourhood_cleansed	TARGET	Los Altos Hills	28	0.7561
TRAIN	neighbourhood_cleansed	TARGET	Morgan Hill	22	0.5941
TRAIN	neighbourhood_cleansed	TARGET	Gilroy	11	0.2971
TRAIN	neighbourhood_cleansed	TARGET	Monte Sereno	6	0.1620

The price for the neighborhood is between $65 to $150 per unit. Los Altos and Los Altos Hills are the highest medians at $150 and the most expensive.

Figure 11

Result of summary statistics neighborhood by price.

Data Role=TRAIN Variable=price

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean
OVERALL		100	0	7221	10	10000	161.2871
neighbourhood_cleansed	Campbell	100	0	131	25	2000	163
neighbourhood_cleansed	Cupertino	99	0	325	24	1200	131.5138
neighbourhood_cleansed	Gilroy	65	0	20	48	385	114.9
neighbourhood_cleansed	Los Altos	150	0	73	30	1985	242.4247
neighbourhood_cleansed	Los Altos Hills	150	0	52	35	2998	297.6538
neighbourhood_cleansed	Los Gatos	100	0	85	36	1000	158.7412
neighbourhood_cleansed	Milpitas	80	0	301	18	799	119.3688
neighbourhood_cleansed	Monte Sereno	125	0	13	85	235	132.3846
neighbourhood_cleansed	Morgan Hill	80	0	42	49	350	108.1667
neighbourhood_cleansed	Mountain View	122	0	664	15	10000	220.887
neighbourhood_cleansed	Palo Alto	134	0	794	10	3000	219.8929
neighbourhood_cleansed	San Jose	87	0	2882	10	3250	134.855
neighbourhood_cleansed	Santa Clara	90	0	711	11	5000	155.1913
neighbourhood_cleansed	Saratoga	110	0	62	39	3400	267.3226
neighbourhood_cleansed	Sunnyvale	95	0	768	15	3000	125.2292
neighbourhood_cleansed	Unincorporated Areas	105	0	298	10	5500	256.349

The result of property type percentage is that the highest rate is House at 51.44%, and the lowest properties are aparthotel, campsite, chalet, and a few more.

Figure 12

Result of summary statistics Property type frequency and percentage.

Data Role=TRAIN

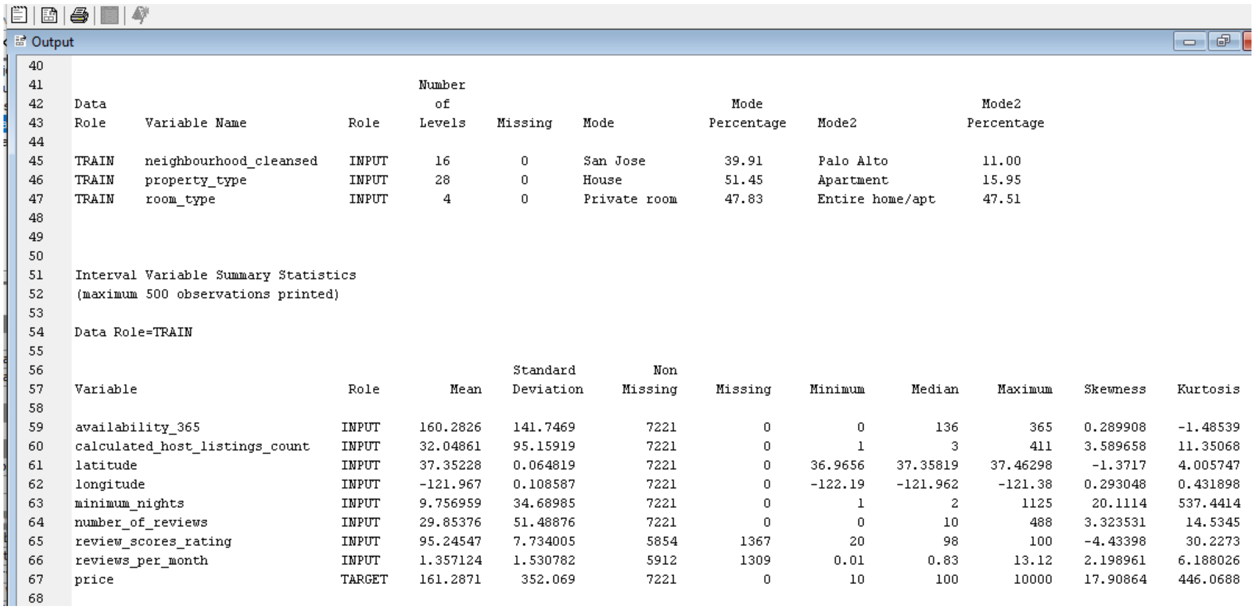
Data Role	Variable Name	Role	Level	Frequency Count	Percent
TRAIN	property_type	TARGET	House	3715	51.4472
TRAIN	property_type	TARGET	Apartment	1152	15.9535
TRAIN	property_type	TARGET	Serviced apartment	504	6.9796
TRAIN	property_type	TARGET	Townhouse	429	5.9410
TRAIN	property_type	TARGET	Guest suite	363	5.0270
TRAIN	property_type	TARGET	Guesthouse	318	4.4038
TRAIN	property_type	TARGET	Condominium	303	4.1961
TRAIN	property_type	TARGET	Villa	129	1.7865
TRAIN	property_type	TARGET	Bungalow	115	1.5926
TRAIN	property_type	TARGET	Loft	41	0.5678
TRAIN	property_type	TARGET	Cottage	29	0.4016
TRAIN	property_type	TARGET	Camper/RV	27	0.3739
TRAIN	property_type	TARGET	Boutique hotel	23	0.3185
TRAIN	property_type	TARGET	Tiny house	16	0.2216
TRAIN	property_type	TARGET	Other	13	0.1800
TRAIN	property_type	TARGET	Bed and breakfast	12	0.1662
TRAIN	property_type	TARGET	Cabin	7	0.0969
TRAIN	property_type	TARGET	Farm stay	6	0.0831
TRAIN	property_type	TARGET	Tent	5	0.0692
TRAIN	property_type	TARGET	Treehouse	3	0.0415
TRAIN	property_type	TARGET	Yurt	3	0.0415
TRAIN	property_type	TARGET	Barn	2	0.0277
TRAIN	property_type	TARGET	Aparthotel	1	0.0138
TRAIN	property_type	TARGET	Campsite	1	0.0138
TRAIN	property_type	TARGET	Chalet	1	0.0138
TRAIN	property_type	TARGET	Earth house	1	0.0138
TRAIN	property_type	TARGET	Lighthouse	1	0.0138
TRAIN	property_type	TARGET	Train	1	0.0138

Data Preparation

The StatExplore node result window shows 'review_scores_rating' has 19% missing data points and ' reviews_per_month' of 18%. In addition, the Graph Explore node result window shows that 'minimum_nights' and 'calculated_host_listings_count' have outliers variables.

Figure 13

Result of summary statistics StatExplore Output Window.



Output

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	neighbourhood_cleansed	INPUT	16	0	San Jose	39.91	Palo Alto	11.00
TRAIN	property_type	INPUT	28	0	House	51.45	Apartment	15.95
TRAIN	room_type	INPUT	4	0	Private room	47.83	Entire home/apt	47.51

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
availability_365	INPUT	160.2826	141.7469	7221	0	0	136	365	0.289908	-1.48539
calculated_host_listings_count	INPUT	32.04861	95.15919	7221	0	1	3	411	3.589658	11.35068
latitude	INPUT	37.35228	0.064819	7221	0	36.9656	37.35819	37.46298	-1.3717	4.005747
longitude	INPUT	-121.967	0.108587	7221	0	-122.19	-121.962	-121.38	0.293048	0.431898
minimum_nights	INPUT	9.756959	34.68985	7221	0	1	2	1125	20.1114	537.4414
number_of_reviews	INPUT	29.85376	51.48876	7221	0	0	10	488	3.323531	14.5345
review_scores_rating	INPUT	95.24547	7.734005	5854	1367	20	98	100	-4.43398	30.2273
reviews_per_month	INPUT	1.357124	1.530782	5912	1309	0.01	0.83	13.12	2.198961	6.188026
price	TARGET	161.2871	352.069	7221	0	10	100	10000	17.90864	446.0688

Figure 14

Frequency Histograms of input variables.

The SAS Output shows large Standard Deviation values of 100 and above: availability_365 of 141.75 and price of 352.069. I will explore transformations to reduce the variance in those variables with large deviations.

Kurtosis shows the variable's probability or frequency, which also helps to compare which variable has a heavy distribution tail with three kurtosis types. All variables have a high kurtosis that is leptokurtic (kurtosis>0), except availability_365, which has close zero and is mesokurtic (kurtosis=0).

Skewness tells whether the dataset has an asymmetric distribution or not that measures three different distributions. Zero skew means the distribution is symmetrical—another negative skew when the number is negative and a positive skew when the number is positive. For example, the skewness of availability_365 shows a close to zero number of 0.29, the symmetrical distribution. Finally, I will look at each continuous variable's distribution to improve the normal distribution of descriptive statistics subjects.

Thus, the availability_365 distribution should have outliers; the skewness result is close to zero, which can be accepted for normal distribution if outliers exceed what we expect.

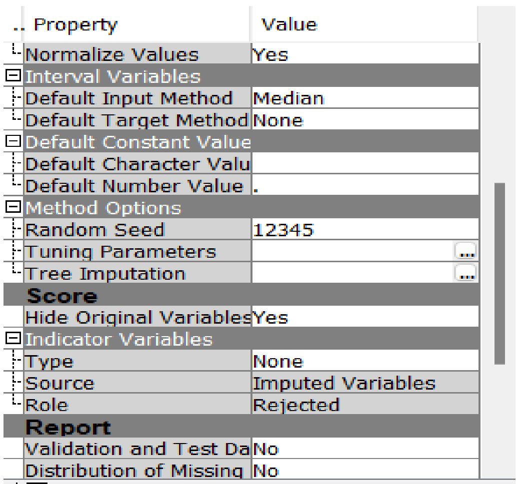
Missing Data

Ca Airbnb data set has missing data points with a high volume for review_scores_rating of 1367 and reviews_per_month of 1309 in a total observation of 7221 that can replace a median number.

Let's look at missing values because missing values can cause the model result. I used the Impute node on the Modify tab for the 'review_scores_rating' and 'reviews_per_month' variables with high missing values. I used the median imputation method for replacing values in skewed distributions.

Figure 15

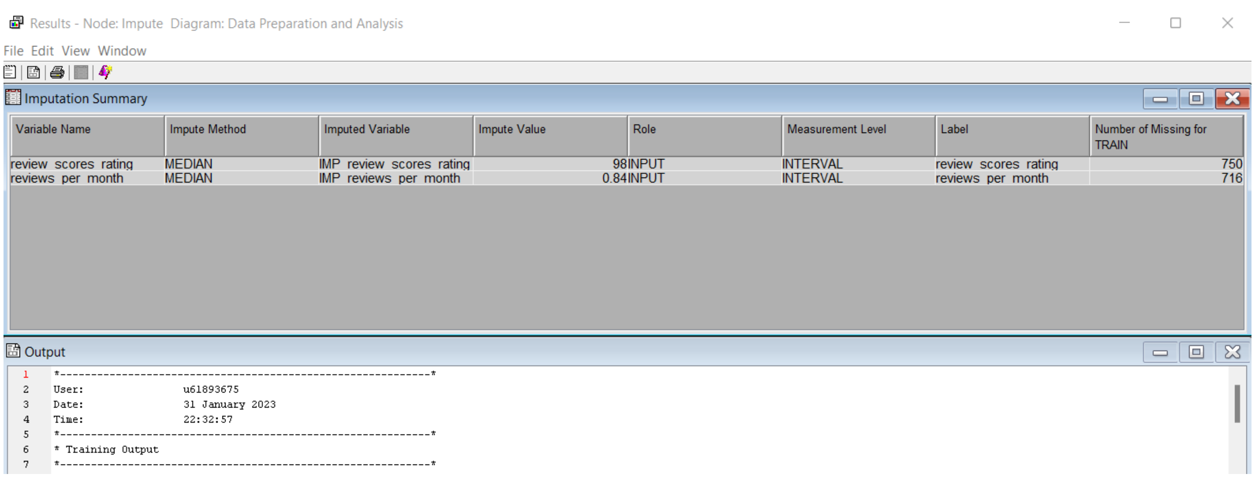
Impute node setting into SAS Enterprise Miner.



Property	Value
Normalize Values	Yes
Interval Variables	
Default Input Method	Median
Default Target Method	None
Default Constant Value	
Default Character Valu	
Default Number Value	.
Method Options	
Random Seed	12345
Tuning Parameters	...
Tree Imputation	...
Score	
Hide Original Variables	Yes
Indicator Variables	
Type	None
Source	Imputed Variables
Role	Rejected
Report	
Validation and Test Da	No
Distribution of Missing	No

Figure 16

Impute Result into SAS Enterprise Miner.



Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
review_scores_rating	MEDIAN	IMP_review_scores_rating	98	INPUT	INTERVAL	review_scores_rating	750
reviews_per_month	MEDIAN	IMP_reviews_per_month	0.84	INPUT	INTERVAL	reviews_per_month	716

Handling Outliers

The Filter node on the Sample tab filters out the outliers. The CA-Airbnb data set variables are ‘minimum_nights,’ and ‘calculated_host_listings_count’ have outlier variables.

The filter setting needs to change under the Train group, Table to Filter set All Data Sets, and under the Interval Variables group, click the Interval Variables' ellipsis and set the minimum for both variables' zero and the maximum value for 'minimum_nights,' variable;

the mean (μ) of 9.756959,

the standard deviation (σ) of 34.68985

as $\mu+\sigma=9.756959+3*34.68985=113.826509$.

The maximum value for the 'calculated_host_listings_count' variable;

the mean (μ) of 32.04861,

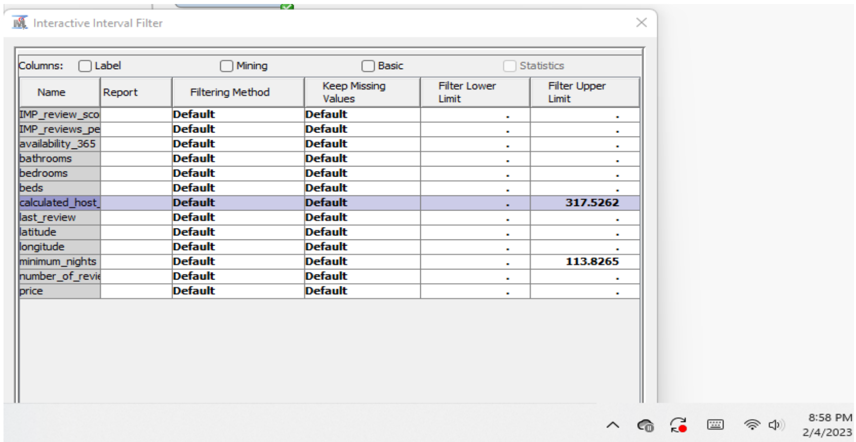
the standard deviation (σ) of 95.15919

as $\mu+\sigma=32.04861+3*95.15919= 317.52618$.

Both values, probably 99.7% of the values, are within three standard deviations. Therefore, the Default Filtering Method is set to User-Specified Limits. Click and update the new Upper Limit value, then click OK, run the Filter node, and view the results.

Figure 17

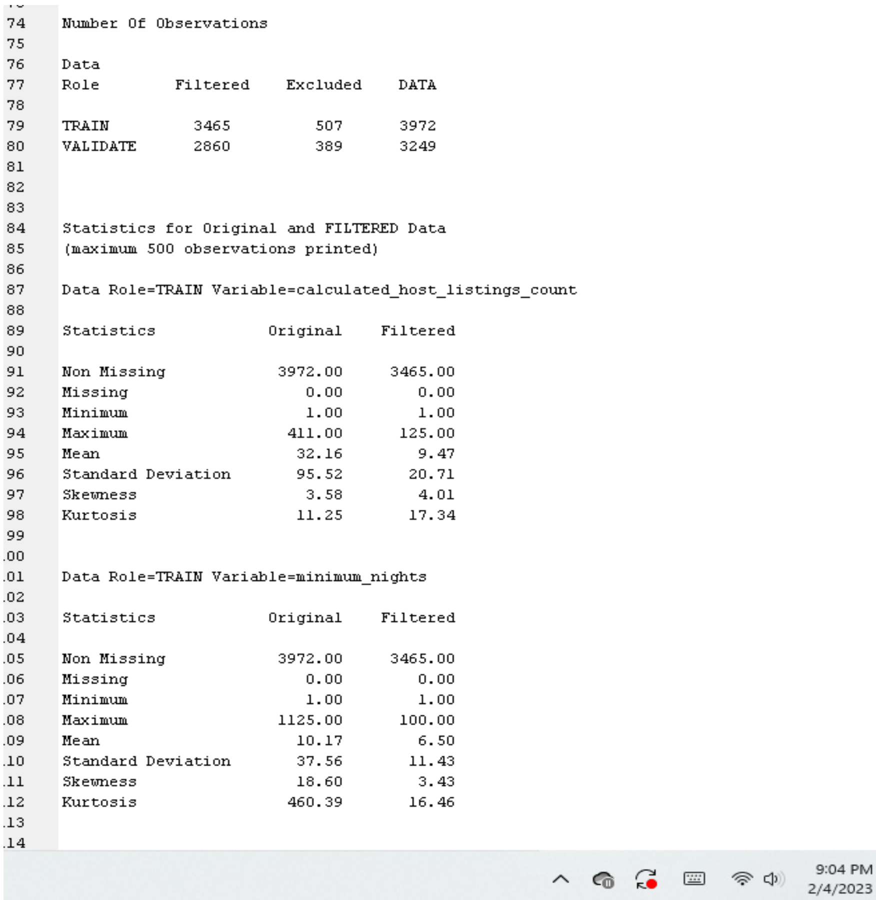
Remove outliers for the Interval variable with the Filter node in SAS Enterprise Mine.



Name	Report	Filtering Method	Keep Missing Values	Filter Lower Limit	Filter Upper Limit
IMP_review_sco		Default	Default	.	.
IMP_reviews_pe		Default	Default	.	.
availability_365		Default	Default	.	.
bathrooms		Default	Default	.	.
bedrooms		Default	Default	.	.
beds		Default	Default	.	.
calculated_host_		Default	Default	.	317.5262
last_review		Default	Default	.	.
latitude		Default	Default	.	.
longitude		Default	Default	.	.
minimum_nights		Default	Default	.	113.8265
number_of_revie		Default	Default	.	.
price		Default	Default	.	.

Figure 18

Result in the Interval variable window.



```
Number Of Observations

Data
Role         Filtered    Excluded    DATA

TRAIN          3465         507      3972
VALIDATE       2860         389      3249



Statistics for Original and FILTERED Data
(maximum 500 observations printed)

Data Role=TRAIN Variable=calculated_host_listings_count

Statistics            Original    Filtered

Non Missing            3972.00     3465.00
Missing                   0.00        0.00
Minimum                   1.00        1.00
Maximum                 411.00      125.00
Mean                     32.16        9.47
Standard Deviation       95.52       20.71
Skewness                  3.58        4.01
Kurtosis                 11.25       17.34


Data Role=TRAIN Variable=minimum_nights

Statistics            Original    Filtered

Non Missing            3972.00     3465.00
Missing                   0.00        0.00
Minimum                   1.00        1.00
Maximum                1125.00      100.00
Mean                     10.17        6.50
Standard Deviation       37.56       11.43
Skewness                 18.60        3.43
Kurtosis                460.39       16.46
```

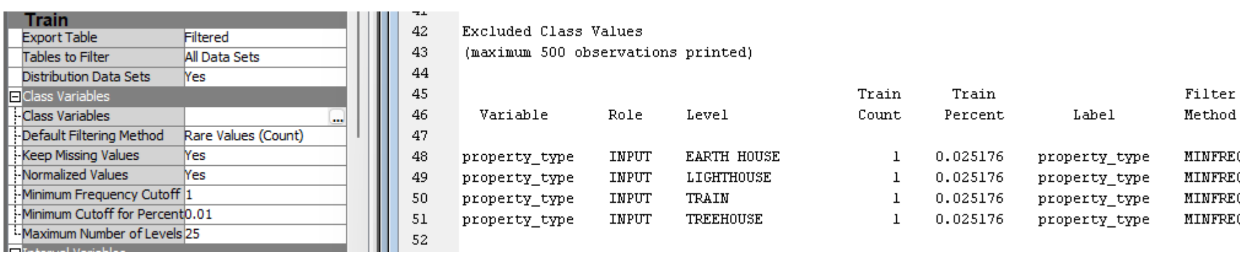
As a result, the train partition has 507, and the validated dataset has 389 filtered observations. The train data display that the maximum 'calculated_host_listings_count' value in the cell is 125, but the original maximum is 411. Likewise, the 'minimum_nights' maximum value in the partition is 100, with a total value of 1125.

Reduce Many Levels of Categorical Variables

One of the topics is that many different levels of the categorical or class variable reduce the performance of the variable. For example, the 'neighborhood_cleansed' has 16 different levels, and 'Property_type' has 28 different level variables. The Replacement node can solve different levels for the groups. For example, I used the Filter node on the Class Variable group to set the 'Property_type' variable default minimum frequency of 25 or can set to manually click ellipsis by Class Variables and the result in the Interactive Class Filter.

Figure 19

Result of Class Variable and properties on Filter node.



CA Airbnb Dataset Descriptive Statistic

Here is the SAS Enterprise Miner StatExplore node result, which describes the CA Airbnb summary statistics as follows:

The most common value or mode for the categorical variable neighborhood area is San Jose, which makes up 41.29%, and the second most frequent value is Palo Alto at 11.71% in For property type, the most frequent variable is House at 53.22%, and the second one is Apartment at 16.96% in Figure 10. Additionally, the variable of room type's most common value is Private room at 47.83%, and the second is Entire home/apt at 47.51% in Figure 7.

Figure 13 shows the result of a summary statistic for interval variables; the mean price amount was $161.2871 per unit with a standard deviation of $352.069 and the median value at $100. The price result of the mean is greater than the median, which means the price dataset has a right skew. The number of reviews' compromise was 29.85376, with a standard deviation of 51.48076, and the median was 10, which is a mean greater than the median as the number of reviews also right skew. The kurtosis of the property's price was 464.0688 as higher than zero means a leptokurtic has a sharper peak than a bell shape. The result of kurtosis for the number of reviews was 14.5345, which means the number of reviews has a leptokurtic kurtosis.

Those variables should be transformed into a normal distribution. I used the Transform Variables node to modify the log transformation for the price and number of reviews. First, review the distribution of the variable by Formulas ellipsis, then the Variables ellipsis set the Log on Method column.

Figure 20

Formulas window- the number of reviews and price of the properties.

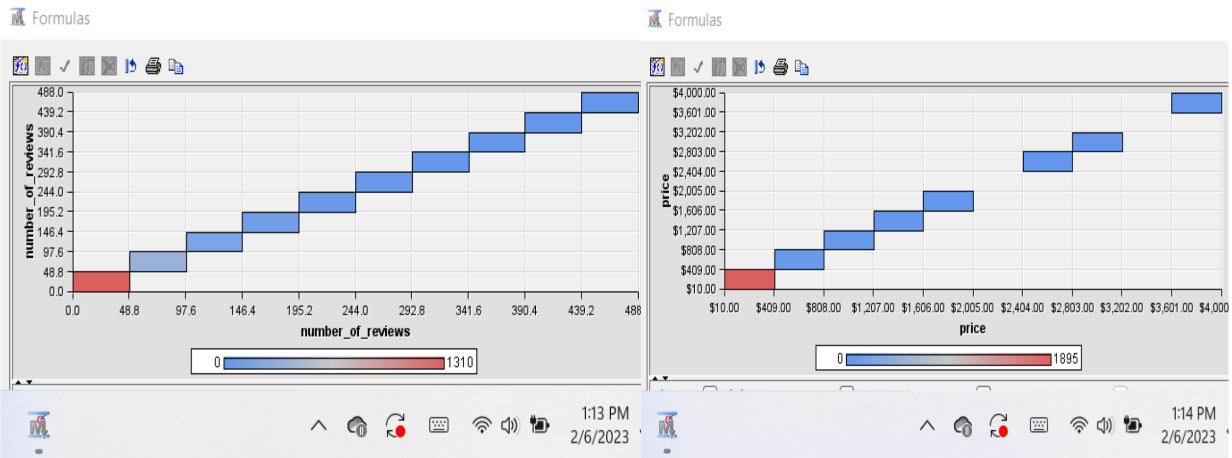
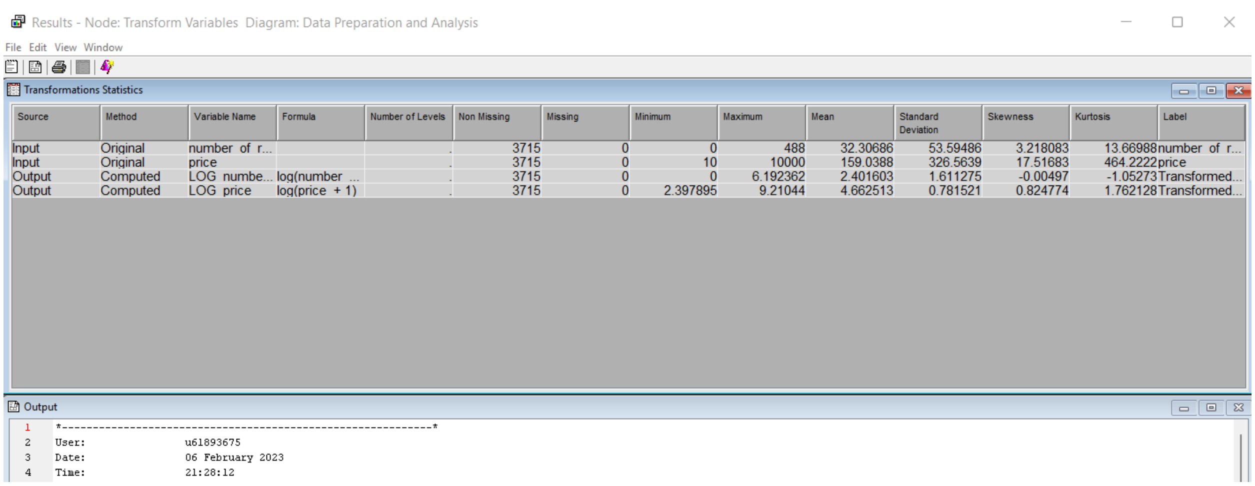


Figure 21

Result of Transform Variables node.



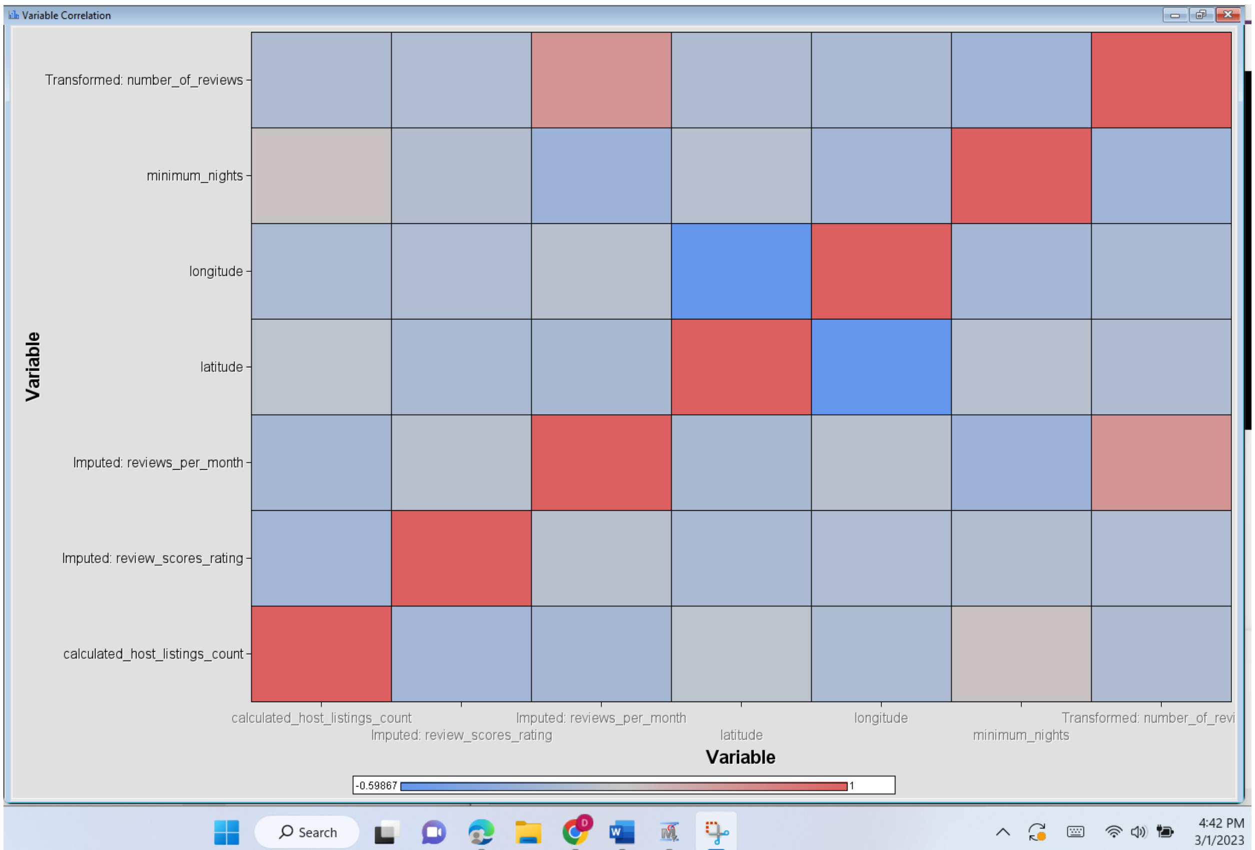
Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Label
Input	Original	number of r...		.	3715	0	0	488	32.30686	53.59486	3.218083	13.66988	number of r...
Input	Original	price		.	3715	0	10	10000	159.0388	326.5639	17.51683	464.2222	price
Output	Computed	LOG numbe...	log(number ...	.	3715	0	0	6.192362	2.401603	1.611275	-0.00497	-1.05273	Transformed...
Output	Computed	LOG price	log(price + 1)	.	3715	0	2.397895	9.21044	4.662513	0.781521	0.824774	1.762128	Transformed...

Covariance and Correlation

One of the other causes of the model's performance is multicollinearity. Correlation analysis helps us to know how to relate two variables, as if the high correlation between two variables may negatively affect the predictive result, as red color is a high relation and blue color is less relation between two input variables. I used the Variable Cluster node in the

Explore tab in SAS Enterprise Miner to find if there is multicollinearity. The CA Airbnb dataset has no collinearity to warrant concerns.

Figure 22: Variable correlation



Business Question and Hypothesis

These are the business problems that I will explore some key points which would be very helpful for business, such as:

1. Does the room type differ based on the property's price?

2. Does the room type differ based on the total number of reviews?

3. Is there a difference in the room types based on the property's availability?

4. Does the neighborhood differ on the property's price?

The organization's strategic goal is that Airbnb's constant goals were to expand into new areas and deliver more inventory within the company's network.

I created the null and alternative hypotheses for each business and the result. Calculating the p-value is complicated, so I will use Chi-Square to find the p-value and compare the null and alternative hypotheses. I set the target value as ‘room_type’ to see each room type’s price mean; the median also helps me to reject the null hypothesis. The StatExplore node includes the Chi-square test.

Business Question 1: Does the room type differ based on the property’s price?

- *Null hypothesis(H10)*: No difference between room types based on property prices exists.
- *Alternative Hypothesis(H1)*: at least one group differs significantly from the overall mean price of the property.

Based on the Figure 24 result, the p-value of the property’s price effect is close to zero and less than the significance level, implying that 0.000 < 0.05. So, we can reject the null hypothesis in favor of an alternative idea. Therefore, in Figure 23, we can conclude that there is a significant difference and that at least one room type differs significantly from the overall mean property price. Simply put, the solution to the business problem indicates that the room types (i.e., Entire home/apt, Private Room, and Shared room) differ for the property’s price.

Figure 23

Result of Chi-square price by room_type with LOG_price.

Data Role=TRAIN Variable=LOG_price

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label
OVERALL		4.564348	0	3719	2.397895	9.21044	4.657673	0.781201	0.840901	1.79121	INPUT	Transformed: price
room_type	Entire home/apt	5.141664	0	1648	2.397895	8.2943	5.203174	0.654337	0.974727	3.032905	INPUT	Transformed: price
room_type	Hotel room	5.141664	0	2	5.141664	5.298317	5.21999	0.110771	.	.	INPUT	Transformed: price
room_type	Private room	4.219508	0	1886	2.397895	9.21044	4.282189	0.539283	2.122231	11.9929	INPUT	Transformed: price
room_type	Shared room	3.433987	0	183	2.944439	7.313887	3.60878	0.561329	3.168122	14.39391	INPUT	Transformed: price

11:53 AM
2/11/2023

Figure 24

Chi-square output for p-value for room type.

Chi-Square Statistics
(maximum 500 observations printed)

Data Role=TRAIN Target=room_type

Input	Chi-Square	Df	Prob
LOG_price	2015.4379	12	<.0001
property_type	1015.5972	57	<.0001
calculated_host_listings_count	364.1993	9	<.0001
neighbourhood_cleansed	177.8311	45	<.0001
longitude	106.2742	12	<.0001
LOG_number_of_reviews	90.6431	12	<.0001
availability_365	82.5775	12	<.0001
IMP_reviews_per_month	82.5441	12	<.0001
minimum_nights	48.1249	12	<.0001
IMP_review_scores_rating	31.8419	12	0.0015
latitude	16.3669	12	0.1750

12:03 PM
2/11/2023

Business Question 2: Does the room type differ based on the total number of reviews?

- *Null hypothesis(H20):* There is no difference between room types based on the total number of reviews.
- *Alternative Hypothesis(H2):* at least one group differs significantly from the overall mean of the total number of reviews.

Based on the Figure 24 result, the p-value of the total number of review effects is close to zero and less than the significance level, implying that 0.000 < 0.05. So, we can reject the null hypothesis in favor of an alternative idea. Therefore, in Figure 25, we can conclude that there is a significant difference and that at least one room type differs significantly from the overall mean of the total number of reviews.

Figure 25

Chi-square result for the number of reviews by room_type with LOG_number_of_review.

Data Role=TRAIN Variable=LOG_number_of_reviews

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label
OVERALL		2.484907	0	3719	0	6.192362	2.397391	1.612524	-0.00081	-1.05482	INPUT	Transformed: number_of_rev
room_type	Entire home/apt	2.833213	0	1648	0	6.100319	2.612663	1.631518	-0.21225	-1.04298	INPUT	Transformed: number_of_rev
room_type	Hotel room	1.791759	0	2	1.791759	4.369448	3.080604	1.822701	.	.	INPUT	Transformed: number_of_rev
room_type	Private room	2.302585	0	1886	0	6.192362	2.251799	1.594067	0.163112	-0.9773	INPUT	Transformed: number_of_rev
room_type	Shared room	2.197225	0	183	0	4.718499	1.95177	1.3603	-0.06886	-1.09061	INPUT	Transformed: number_of_rev

12:13 PM 2/11/2023

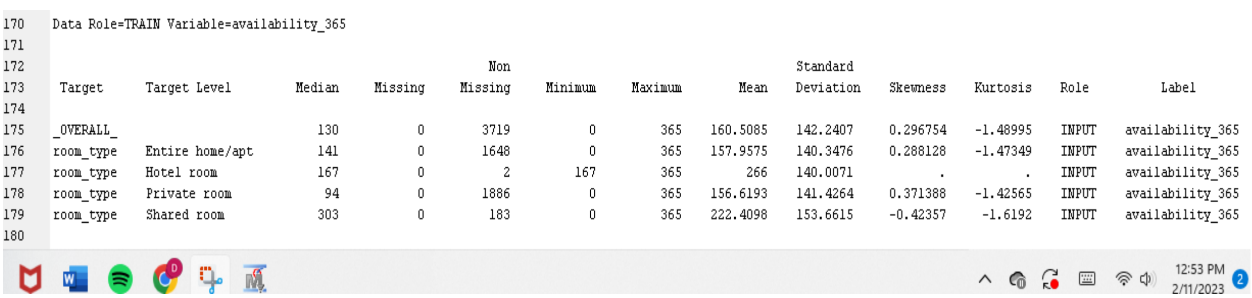
Business Question 3: Is there a difference in the room types based on the property's availability?

- *Null hypothesis(H30):* There is no difference between room types based on the property's availability.
- *Alternative Hypothesis(H3):* at least one group differs significantly from the overall mean of property availability.

Based on the Figure 24 result, the p-value of the room type effect is close to zero and less than the significance level, implying that 0.000 < 0.05. So, we can reject the null hypothesis in favor of an alternative idea. Therefore, in Figure 26, we can conclude that there is a significant difference and that at least one room type differs significantly from the overall mean of availability of the property.

Figure 26

Result of Chi-square test for property type by room_type with availablety_365.



Data Role=TRAIN Variable=availability_365

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label
OVERALL		130	0	3719	0	365	160.5085	142.2407	0.296754	-1.48995	INPUT	availability_365
room_type	Entire home/apt	141	0	1648	0	365	157.9575	140.3476	0.288128	-1.47349	INPUT	availability_365
room_type	Hotel room	167	0	2	167	365	266	140.0071	.	.	INPUT	availability_365
room_type	Private room	94	0	1886	0	365	156.6193	141.4264	0.371388	-1.42565	INPUT	availability_365
room_type	Shared room	303	0	183	0	365	222.4098	153.6615	-0.42357	-1.6192	INPUT	availability_365

Business Question 4: Does the neighborhood differ on the property's price?

- *Null hypothesis(H40):* There is no difference between the neighborhood of the property based on the price of the property.
- *Alternative Hypothesis(H4):* at least one group differs significantly from the overall mean cost of the property.

Based on the Figure 27 result, the p-value of the effect is close to zero and less than the significance level, implying that 0.002 < 0.05. So, we can reject the null hypothesis in favor of an alternative idea. Therefore, in Figure 28, we can conclude that there is a significant difference and that at least one neighborhood group differs significantly from the overall mean property price.

Figure 27

Chi-Square statistics result of p-value by neigborhood_cleansed dataset.

Chi-Square Statistics
(maximum 500 observations printed)

Data Role=TRAIN Target=neighbourhood_cleansed

Input	Chi-Square	Df	Prob
latitude	6709.1199	60	<.0001
longitude	5903.5813	60	<.0001
calculated_host_listings_count	320.2544	45	<.0001
room_type	159.1430	30	<.0001
property_type	952.7695	285	<.0001
LOG_price	222.5418	60	<.0001
minimum_nights	156.5336	60	<.0001
availability_365	141.1279	60	<.0001
IMP_review_scores_rating	108.1250	60	0.0001
IMP_reviews_per_month	106.1387	60	0.0002
LOG_number_of_reviews	95.2241	60	0.0026

```
*------------------------------------------------------------*
* Score Output
*------------------------------------------------------------*
```

12:33 PM
2/11/2023

Figure 28

Chi-square results The HP Neural node is also priced by neighborhood properties.

Data Role=TRAIN Variable=LOG_price

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	
OVERALL		4.564348	0	3703	2.397895	9.21044	4.658408	0.777507	0.819146	1.757767	INPUT	Tra
neighbourhood_cleansed	Campbell	4.615121	0	63	3.583519	7.601402	4.78668	0.703812	1.564041	4.259307	INPUT	Tra
neighbourhood_cleansed	Cupertino	4.60517	0	154	3.367296	7.09091	4.66171	0.659701	0.501152	0.220483	INPUT	Tra
neighbourhood_cleansed	Gilroy	4.189655	0	11	3.912023	5.955837	4.56025	0.769149	1.183742	-0.13488	INPUT	Tra
neighbourhood_cleansed	Los Altos	5.01728	0	39	3.433987	7.593878	5.115349	0.893735	0.555657	0.53253	INPUT	Tra
neighbourhood_cleansed	Los Altos Hills	5.01728	0	28	3.713572	6.857514	4.981886	0.884821	0.230525	-0.65051	INPUT	Tra
neighbourhood_cleansed	Los Gatos	4.564348	0	44	3.663562	6.908755	4.734907	0.701597	1.440269	2.364813	INPUT	Tra
neighbourhood_cleansed	Milpitas	4.369448	0	155	3.258097	6.356108	4.459868	0.75067	0.338211	-0.94812	INPUT	Tra
neighbourhood_cleansed	Monte Sereno	4.867534	0	6	4.454347	5.170484	4.828839	0.238101	-0.31233	1.071792	INPUT	Tra
neighbourhood_cleansed	Morgan Hill	4.26268	0	22	3.912023	5.141664	4.364835	0.378746	0.718409	-0.30052	INPUT	Tra
neighbourhood_cleansed	Mountain View	4.70048	0	326	2.772589	9.21044	4.765927	0.748758	1.524745	7.310179	INPUT	Tra
neighbourhood_cleansed	Palo Alto	4.912655	0	423	2.397895	8.006701	5.036578	0.787957	0.954081	1.851918	INPUT	Tra
neighbourhood_cleansed	San Jose	4.454347	0	1529	2.397895	7.601402	4.565328	0.752013	0.605181	0.465756	INPUT	Tra
neighbourhood_cleansed	Santa Clara	4.394449	0	340	2.484907	7.824446	4.563231	0.778202	0.624173	1.061466	INPUT	Tra
neighbourhood_cleansed	Saratoga	4.564348	0	32	3.931826	7.409136	4.846208	0.769532	1.55542	2.823123	INPUT	Tra
neighbourhood_cleansed	Sunnyvale	4.394449	0	376	2.772589	8.006701	4.510398	0.640257	0.700701	2.33138	INPUT	Tra
neighbourhood_cleansed	Unincorporated Areas	4.615121	0	155	2.397895	8.2943	4.837514	1.052758	1.028851	1.264055	INPUT	Tra

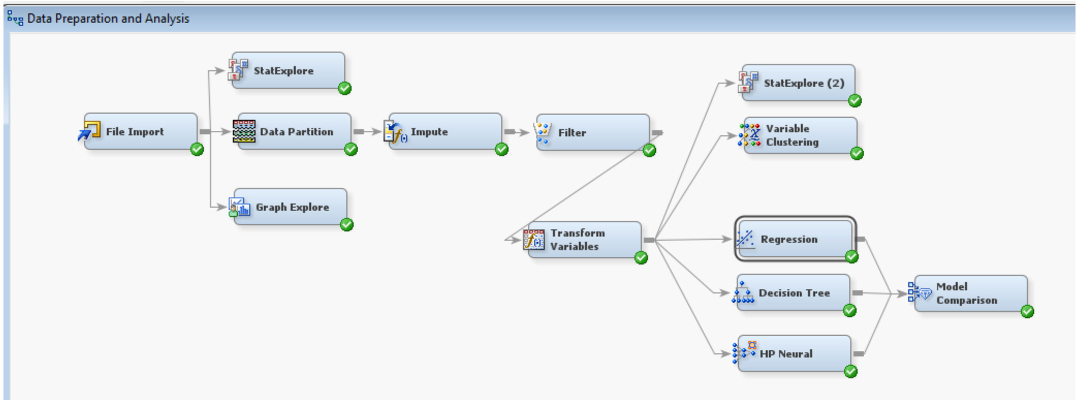
12:34 PM
2/11/2023

Predictive Analysis

After achieving descriptive statistics and data preparation, the next step is building the predictive model. I will then use the Big Three models: linear regression, Decision tree, and HP Neural network.

Figure 29

Node process flow of models.

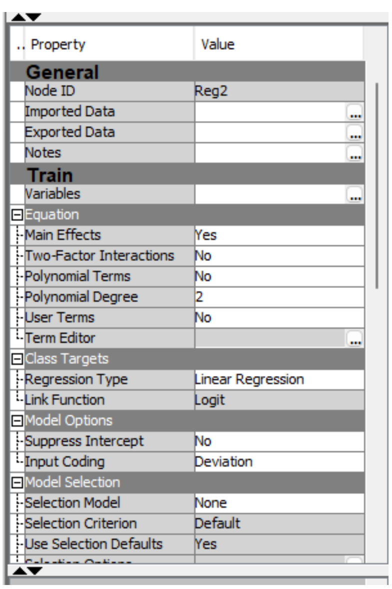


Linear Regression Model

The most popular model is linear regression, which allows us to measure the strength of the relationship between the response and predictor variable, also known as line fitting and curve fitting. The case target variable is ‘price,’ which is the continuous variable, so I apply a linear regression model.

Figure 30

Linear regression properties.



Property	Value
General	
Node ID	Reg2
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Linear Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	None
Selection Criterion	Default
Use Selection Defaults	Yes

The average square error of 0.279806 on the Validation partition. Suppose the high difference between the Train and Validation partition is the average squared error is insignificant. In that case, the result of the difference is 0.003375 for the claim fraud train and validation partition. The model does not appear to be overfitting.

Figure 31

Fit Statistics of Linear Regression result.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	T
LOG_price	Transformed:...	AIC	Akaike's Infor...	-4597.1	.	
LOG_price	Transformed:...	ASE	Average Squ...	0.283181	0.279806	
LOG_price	Transformed:...	AVERR	Average Erro...	0.283181	0.279806	
LOG_price	Transformed:...	DFE	Degrees of F...	3670	.	
LOG_price	Transformed:...	DFM	Model Degre...	45	.	
LOG_price	Transformed:...	DFT	Total Degree...	3715	.	
LOG_price	Transformed:...	DIV	Divisor for ASE	3715	3051	
LOG_price	Transformed:...	ERR	Error Function	1052.019	853.6883	
LOG_price	Transformed:...	FPE	Final Predicti...	0.290126	.	
LOG_price	Transformed:...	MAX	Maximum Ab...	3.831268	3.633768	
LOG_price	Transformed:...	MSE	Mean Square...	0.286654	0.279806	
LOG_price	Transformed:...	NOBS	Sum of Frequ...	3715	3051	
LOG_price	Transformed:...	NW	Number of E...	45	.	
LOG_price	Transformed:...	RASE	Root Average...	0.532148	0.528967	
LOG_price	Transformed:...	RFPE	Root Final Pr...	0.538633	.	
LOG_price	Transformed:...	RMSE	Root Mean S...	0.5354	0.528967	
LOG_price	Transformed:...	SBC	Schwarz's B...	-4317.19	.	
LOG_price	Transformed:...	SSE	Sum of Squa...	1052.019	853.6883	
LOG_price	Transformed:...	SUMW	Sum of Case...	3715	3051	

12:55 PM
3/6/2023

The most significant result is close to zero, and Analysis of Variance shows model statistical information like the F value measures how a group of variables is jointly substantial. The P value shows how strongly this model supported the CA-Airbnb dataset. Thus, Pr> F is <0.001, which means the model supports the CA-Airbnb dataset.

The Type 3 Analysis of Effect table contains each variable Pr>F value; if the result is close to 1, the insignificant input variables can be removed as the input variable if statistically significant input should be included in further analysis. The most variable is statistically significant except 'latitude' and 'longitude.'

Figure 32

Output window of Linear Regression Model.

```
                              Analysis of Variance

                                      Sum of
Source                   DF          Squares     Mean Square    F Value    Pr > F

Model                    44      1216.400611       27.645468      96.44    <.0001
Error                  3670      1052.018874        0.286654
Corrected Total        3714      2268.419485


              Model Fit Statistics

R-Square         0.5362     Adj R-Sq         0.5307
AIC          -4597.0952     BIC          -4593.9920
SBC          -4317.1892     C(p)            45.0000


                        Type 3 Analysis of Effects

                                                 Sum of
Effect                             DF           Squares     F Value    Pr > F

IMP_review_scores_rating            1            7.9737       27.82    <.0001
IMP_reviews_per_month               1            2.0889        7.29    0.0070
LOG_number_of_reviews               1           23.8399       83.17    <.0001
availability_365                    1           13.1203       45.77    <.0001
calculated_host_listings_count      1            5.4873       19.14    <.0001
latitude                            1            0.1544        0.54    0.4631
longitude                           0             .             .         .
minimum_nights                      1            6.4865       22.63    <.0001
neighbourhood_cleansed             15           49.7926       11.58    <.0001
property_type                      19          113.6039       20.86    <.0001
room_type                           3          879.6795     1022.93    <.0001
```

1:27 PM
3/6/2023

Decision Tree

Decision trees are the most popular predictive and descriptive-analytic because it is easy to create and understand at least one categorical or continuous target variable as I applied the CA-Airbnb dataset with the price as a constant target variable. Criteria used for evaluating performance will be an average squared error, lift charts, and misclassification rates.

Figure 33

Properties of Decision Tree Model.

Property	Value
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Average Square Error
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Importa	
Observation Based Importa	No
Number Single Var Importar	5
P-Value Adjustment	
Bonferroni Adjustment	Yes
Time of Bonferroni Adjustm	Before

Assessment Measure

The result of the average square error at 0.264227 by validation dataset that statistical result helps to compare the predictive model. The previous mode is the linear regression model result of an average square error of 0.279806 on the Validation partition. The decision tree model is slightly better than the linear regression because of the lower average square error.

Figure 34

Fit Statistics window of Decision Tree.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
LOG price	Transfor...	NOBS	Sum of Fr...	3715	3051	.
LOG price	Transfor...	MAX	Maximum...	4.916402	4.582786	.
LOG price	Transfor...	SSE	Sum of S...	892.8402	806.1565	.
LOG price	Transfor...	ASE	Average ...	0.240334	0.264227	.
LOG price	Transfor...	RASE	Root Aver...	0.490239	0.51403	.
LOG price	Transfor...	DIV	Divisor fo...	3715	3051	.
LOG price	Transfor...	DFT	Total Deg...	3715	.	.

2:29 PM
3/6/2023

The Variable Importance window shows a list of input variables used in the decision tree and the number of splits obtained within those variables. The importance of statistics for the training dataset shows how the input variables fit the tree. The decision tree Variable Importance result shows that the ten input variables are the ratio of validation importance. The room_type affects the entire tree, and property_type is the second variable that affects the tree.

Figure 35

Variable Importance of Decision Tree.

File Edit View Window

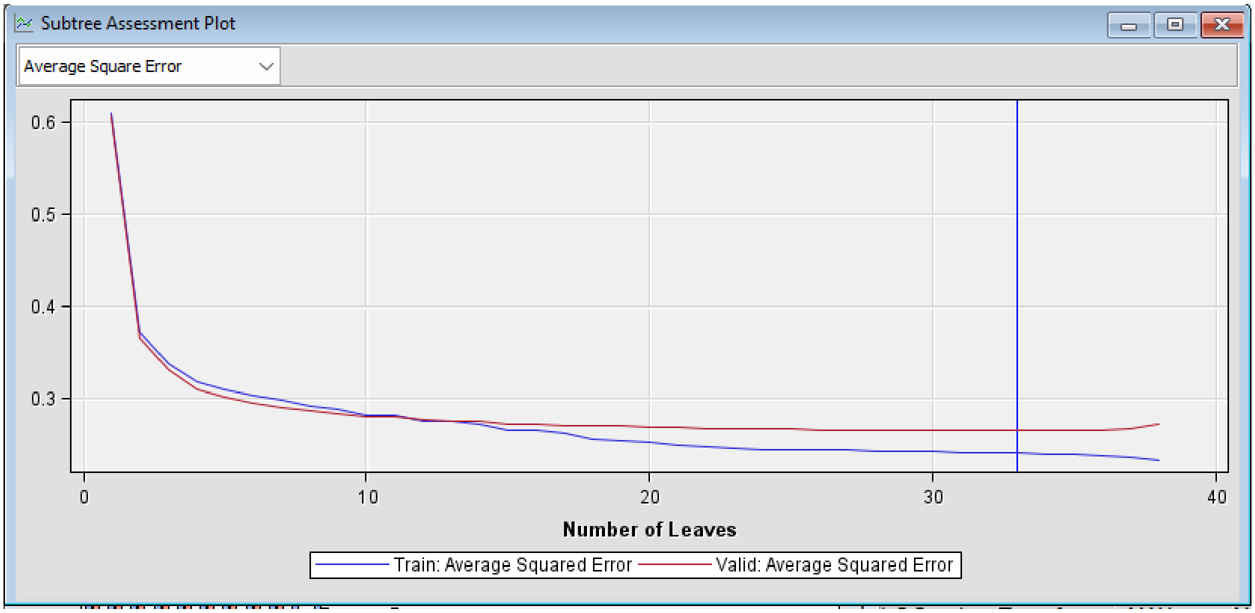
Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
room type	room type	2	1.0000	1.0000	1.0000
property type	property type	2	0.3920	0.3880	0.9898
calculated host listings c...	calculated host listings c...	7	0.2664	0.1580	0.5930
LOG number of reviews	Transformed: number of r...	3	0.2609	0.2464	0.9443
longitude	longitude	6	0.2557	0.1825	0.7140
availability 365	availability 365	3	0.1768	0.2109	1.1929
minimum nights	minimum nights	2	0.1222	0.0145	0.1188
neighbourhood cleansed	neighbourhood cleansed	2	0.1038	0.0522	0.5027
IMP reviews per month	Imputed: reviews per month	2	0.1014	0.0321	0.3162
latitude	latitude	3	0.0796	0.0557	0.7000
IMP review scores rating	Imputed: review scores ra...	0	0.0000	0.0000	.

The Subtree Assessment plot helps to know if the model has been overfitting. If it fits the data, it becomes less comprehensive; the resulting decision tree model is balanced.

Figure 36

Subtree Assessment Plot of Decision Tree.



Neural Network Model with HP Neural Node

The HP Neural node creates a neural network that best applies a large amount of data to make it practical for the most significant decision-making. The HP Neural node can process with interval, binary, or nominal target variable(s) and input(s). Additionally, the HP Neural node eliminates significant data movement and builds advances in parallel processing and in-line memory.

Figure 37

HP Neural model properties.

Property	Value
Network Options	
Input Standardization	Range
Architecture	One Layer
Number of Hidden Neurons	3
Number of Hidden Layers	3
Hidden Layer Options	...
Direct Connections	No
Target Standardization	Range
Target Activation Function	Identity
Target Error Function	Normal
Number of Tries	2
Maximum Iterations	300
Use Missing as Level	No
Report	
Maximum Number of Links	1000
Status	
Create Time	3/3/23 8:13 PM
Run ID	86a567ea-6ae7-7742-829a
Last Error	
Last Status	Complete
Last Run Time	3/6/23 8:15 PM

The Output window contains model information as the Limited Memory BFGS algorithm processed; three hidden neurons and one hidden layer were created, 163 weights were used, 6766 observations were read, and only 3715 were used to train the model.

Figure 38

Output of HP Neural model.

Output

```
Data                   Engine    Role     Path

WORK.HPNNA_TRAINDATA    V9       Input    On Client


                Model Information

Data Source                  WORK.HPNNA_TRAINDATA
Architecture                 MLP
Number of Input Variables    11
Number of Hidden Layers      1
Number of Hidden Neurons     3
Number of Target Variables   1
Number of Weights            163
Optimization Technique       Limited Memory BFGS


Number of Observations Read           6766
Number of Observations Used           6766
Number Used for Training              3715
Number Used for Validation            3051
```

The result of the average square error at 0.240371 by validation dataset. The previous mode is the decision tree model result of an average square error at 0.264227 on the validation dataset. The neural network model is slightly better than the decision tree because of the lower average square error.

Figure 39

Fit Statistics result of HP Neural.

Fit Statistics

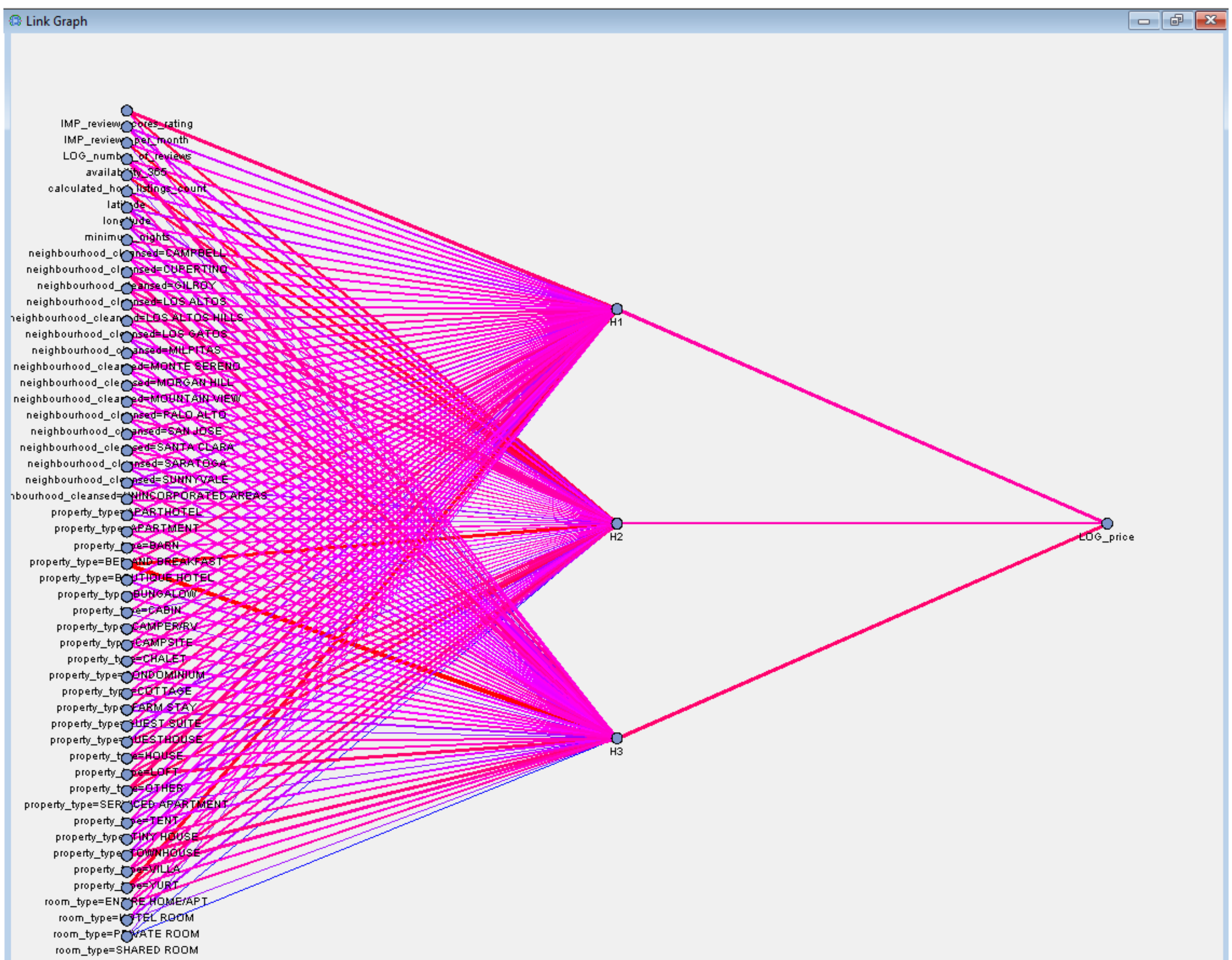
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
LOG_price	Transfor...	ASE	Average ...	0.240371	0.246757	.
LOG_price	Transfor...	DIV	Divisor fo...	3715	3051	.
LOG_price	Transfor...	MAX	Maximum...	3.673405	4.039275	.
LOG_price	Transfor...	NOBS	Sum of Fr...	3715	3051	.
LOG_price	Transfor...	RASE	Root Aver...	0.490276	0.496747	.
LOG_price	Transfor...	SSE	Sum of S...	892.977	752.8563	.

In the HPNeural node, one of the results is a Link Graph, which the neural network doesn't. The Link Graph plot is an ideal representation of the neural network. The target variable is always on the right side, and the input(s) are on the left. A color difference on the line means the thicker red line represents a more significant link weight, and the thinner blue represents a smaller one.

Thus, in this neural network, extensive weight links are room_type=Entire Home property_type-Boutique Hotel, Log_number_of_reviews.

Figure 40

Link Graph of HP Neural model.



Model Comparison

After the 3 Big models' results, the next step is to compare which best fits the CA_Airbnb dataset. The Model Comparison node compares models and predictions from other models and then decides on the best-fit model for the CA-Airbnb dataset to help us find statistically significant predictors for price decisions. I will compare linear regression, decision tree, and HP neural.

The result of output Fit statistics tables contains each model ASE as the selected Model column shows Y which HP Neural model is the best fit.

Figure 41

Fit Statistics by Model Comparison Node.

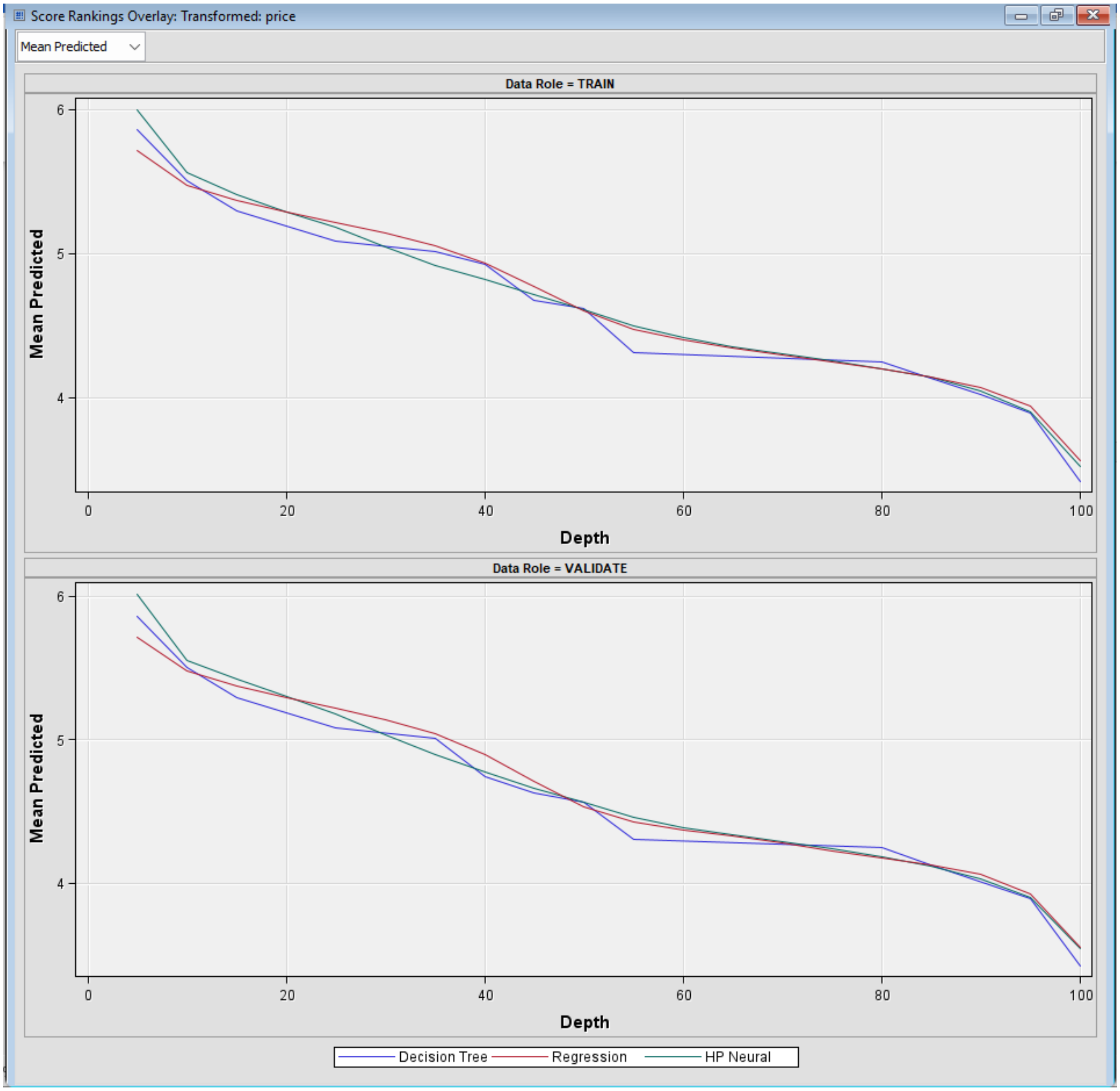
```
Fit Statistics
Model Selection based on Valid: Average Squared Error (_VASE_)

                                                Valid:      Train:
                                                Average     Average
Selected      Model         Model               Squared     Squared
 Model        Node       Description             Error       Error

   Y          HPNNA      HP Neural              0.24676     0.24037
              Tree       Decision Tree          0.26423     0.24033
              Reg2       Regression             0.27981     0.28318
```

11:46 AM
3/8/2023

The Score Ranking Overlay with Mean Predicted selection shows that each model predicts mean price. The x-axis shows the result of the mean predicted, and the y-axis shows the depth, which is how far the model is from random guessing. The higher the mean predicted, the better. Thus, the highest mean predicted was 5.99 and 5% depth from the HP Neural model on the validation dataset.

Figure 42

Score Rankings Overlay by Model Comparison node.



Conclusion

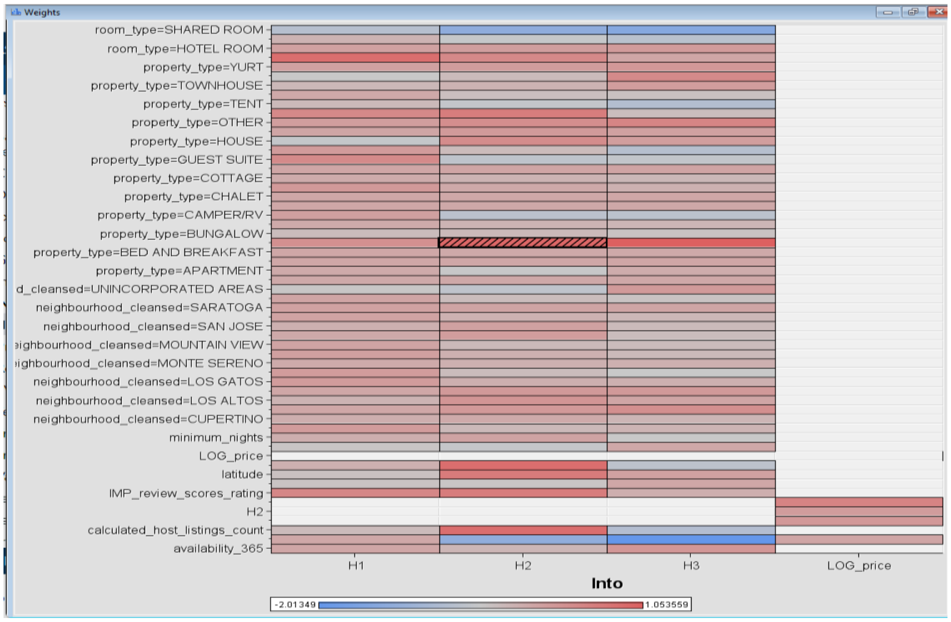
My project involved exploring the CA-Airbnb.csv dataset to gain insight into the price predicted and which variables are predicted. I started by discussing and dividing the dataset into training and validation partitions. I then used linear regression, decision trees, and neural networks to train my models on the training set and evaluate their performance on the validation set. After comparing the models' results, I concluded that the HP Neural model was the most effective for the CA-Airbnb.csv dataset. I also used the Weights graph by the HP Neural model to explore the data further and gain insights into the results.

Model Name	Average Square error	Depth	Mean Predicted
Linear Regression	0.279808	5%	5.71
Decision Tree	0.264227	5%	5.85
HP Neural	0.246757	5%	5.99

The result of HP Neural's essential variables for the Weight window contains color; the dark red heavy weight means the most significant link weight and the blue color values indicate a smaller link weight. Thus, the CA-Airbnb dataset's most predictive variable for the target variable is price; room_type= Entire home/Apt at 0.85, property type=Boutique Hotel at 1.05, and number of reviews at 0.84.

Figure 43

Weight window from HP Neural.



-A high percentage of room types by neighborhood, as hotel rooms in Palo Alto and Santa Clara are 50%.

- The room type Entire Home/Apt price median is $168 per unit, and the Private room price is the median at $67 per unit.

- The highest number of room type private room reviews of 6.19.

- The highest mean availability by room type is a hotel room 266.

Overall, the predictive project was a great learning experience. I learned about the various data predictive techniques and how they can be used to gain insights from the data. I also learned about the different methods that can be used to evaluate the performance of the models. Additionally, I gained a better understanding of the importance of data visualization and how it can be used to identify patterns and relationships in the data.

References

Richard V. McCarthy; Mary M. McCarthy; Wendy Ceccucci, 2022. *Applying Predictive Analytics Finding Value in Data*. Second edition.

Bode, O., Toader, V., & Rus, R. (2022). Pricing Strategies of Porto's Airbnb New Listings. International Conference On Tourism Research, 15(1), 425-432. https://doi.org/10.34190/ictr.15.1.249

Guggilla, Chakraborty, P. Price Recommendation Engine for Airbnb. Support.sas.com. Retrieved from https://support.sas.com/resources/papers/proceedings17/1326-2017.pdf

Kas, J., Delnoij, J., Corten, R., & Parigi, P. (2022). Trust spillovers in the sharing economy: Does international Airbnb experience foster cross-national trust? Journal Of Consumer Behaviour, 21(3), 509-522. https://doi.org/10.1002/cb.2014

SAS Documentation. Support.sas.com. (2022). Retrieved from https://support.sas.com/en/documentation.html

Vlogger, M., Pforr, C., Stawinoga, A., Taplin, R., & Matthews, S. (2018). Who adopts the Airbnb innovation? An analysis of international visitors to Western Australia. Tourism Recreation Research, 43(3), 305-320. https://doi.org/10.1080/02508281.2018.1443052

Ray William, 2020. Bay Area, CA – Airbnb Data (UPDATED 2020). https://www.kaggle.com/datasets/raywilliam/bay-area-airbnb-data-updated-2020

Way X, 2020. New York City Airbnb Data Analysis, Visualization, and Prediction9V2). https://tyuion1215.medium.com/newyorkcity-airbnb-data-analysis-visualization-and-predication-8397943066f9

PropertyManagement, 2022. Airbnb Statistics. https://ipropertymanagement.com/research/airbnb-statistics

J.Li, F. Biljecki.(2019). The Implementation of Big Data Analysis in Regulating Online Short-Term Rental Business: A Case of Airbnb in Beijing. https://ual.sg/publication/2019-sdsc-airbnb-beijing/2019-sdsc-airbnb-beijing.pdf