HOTEL BOOKING DEMANDS

PREDICTIVE ANALYTICS MODELS

BOOKINGS CANCELLATION PREDICTIONS: REVENUE MANAGEMENT

FINAL PROJECT DSCI 5420: BY GROUP 30

EXECUTIVE SUMMARY

The global hotel industry is a multibillion dollar industry. Using the Hotel Booking Demand data set gathered we can gain insight on two hotels, Resort and City, and the possible decisions made behind the scenes. Using different statistical models, such as **Regression Analysis**, **Decision Tree**, and **Neural Network** models, we are focusing on extrapolate trends to add value to the hotels in the data.

All the models were prepped as similarly as possible to obtain the most cohesive results. It was discovered that the comparison between the models can be interpreted differently depending on the static variable used. For example using the Average squared error versus Mean squared error provided us with varying analysis. Ultimately, we used the variable that gave the group the most meaningful analysis: the Decision Tree, followed by Logistic Regression.

These two models revealed that variables such as customer lead times and deposits had high significance on cancelled reservations.

To combat this, our group recommends that both hotels implement non-refundable deposits for longer lead time reservations in order to recuperate costs spent on forecasting staff and preparations.

PROJECT BACKGROUND

The foundation of this project is based on the use of second-hand data which was gathered from the website Kaggle. The data, however, originated from the article “Hotel Booking Demand datasets” written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019. The data in Kaggle is titled “Hotel Bookings” and it is real data with false customer identification to protect the privacy and safety of the guests. The four data columns, 'name', 'email', 'phone number' and 'credit_card' have been artificially created and added to the dataset.This dataset was gathered between the 1st of July 2015 until the 31st of August of 2017. It contains 36 variables and 119390 observations from two hotels, a city hotel and a resort hotel.

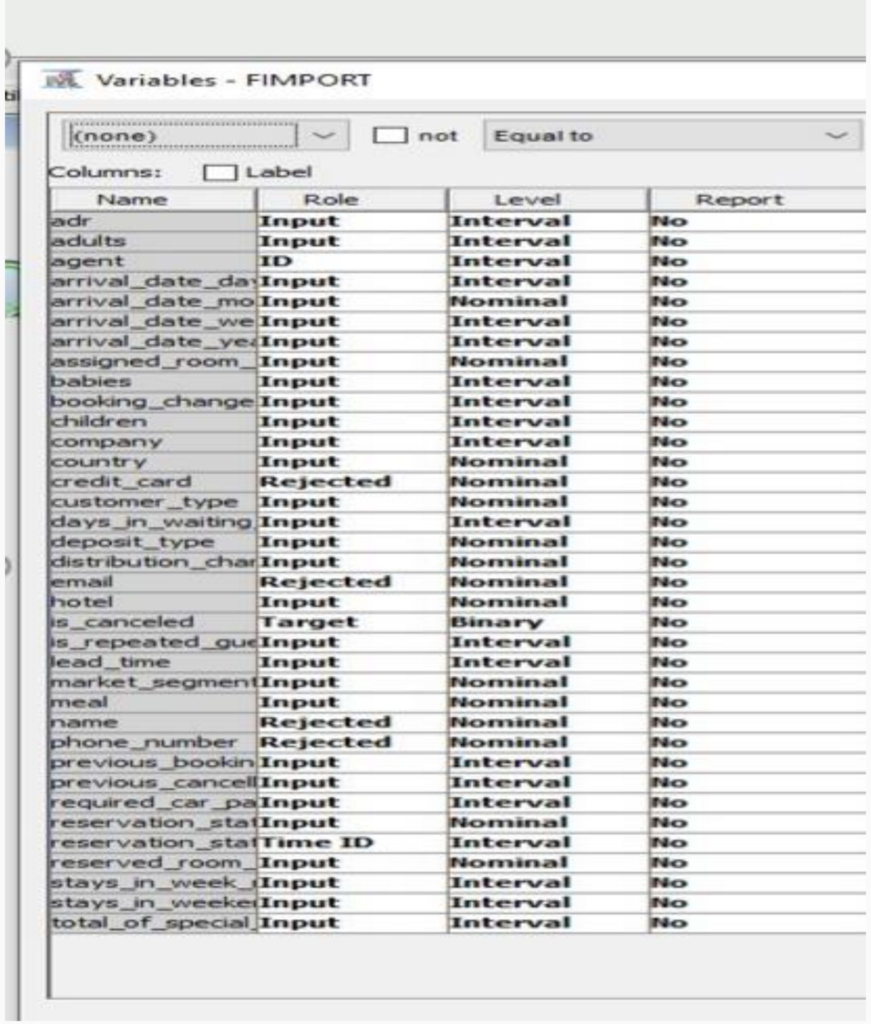
This data set was chosen over many others due to its analysis potential. Using this data set, our group hopes to provide analysis using statistical models to deduce business solutions to add value to the hotels being observed.

DATA DESCRIPTION

The data set contains 36 columns and 119390 rows of nominal, interval, and binary data.

During the roles assignment of the variables, four variables were originally rejected due to falsification. These variables were artificially created to provide a user identity while protection for customers. These were 'name', 'email', 'phone number' and 'credit_card.

Roles were also separated in Input, Time_ID, Rejected, ID, and Target variables.



Variables - FIMPORT

(none) not Equal to

Columns: Label

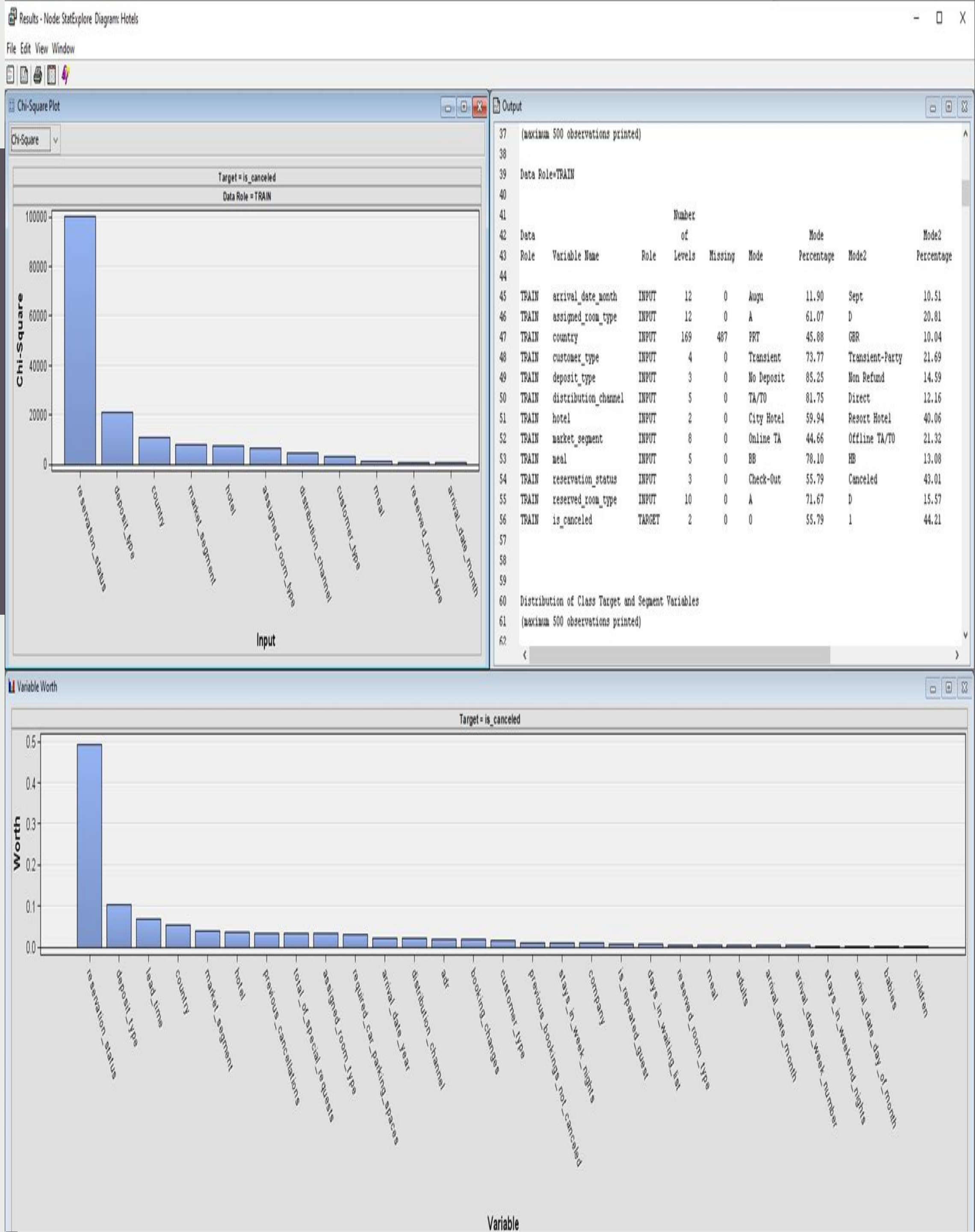
Name	Role	Level	Report
adr	Input	Interval	No
adults	Input	Interval	No
agent	ID	Interval	No
arrival_date_da	Input	Interval	No
arrival_date_mo	Input	Nominal	No
arrival_date_we	Input	Interval	No
arrival_date_ye	Input	Interval	No
assigned_room_	Input	Nominal	No
babies	Input	Interval	No
booking_change	Input	Interval	No
children	Input	Interval	No
company	Input	Interval	No
country	Input	Nominal	No
credit_card	Rejected	Nominal	No
customer_type	Input	Nominal	No
days_in_waiting	Input	Interval	No
deposit_type	Input	Nominal	No
distribution_char	Input	Nominal	No
email	Rejected	Nominal	No
hotel	Input	Nominal	No
is_canceled	Target	Binary	No
is_repeated_gue	Input	Interval	No
lead_time	Input	Interval	No
market_segment	Input	Nominal	No
meal	Input	Nominal	No
name	Rejected	Nominal	No
phone_number	Rejected	Nominal	No
previous_bookin	Input	Interval	No
previous_cancell	Input	Interval	No
required_car_pa	Input	Interval	No
reservation_stat	Input	Nominal	No
reservation_stat	Time ID	Interval	No
reserved_room_	Input	Nominal	No
stays_in_week_	Input	Interval	No
stays_in_weeke	Input	Interval	No
total_of_special_	Input	Interval	No

DATA PREPARATION ACTIVITIES

After roles were determined, the Stat Explore node was used to view the data in bar chart form for variable average width as well as to see the number of missing variables.

Variables such as ‘country’ and ‘agent’ later were determined to have too many missing variables and to be too difficult to run. After filtering and imputing the data the group decided to reject these variables.

The data set was then partitioned and imputed to remove missing variables that would harm the results of the models.



MODELS USED

Our group chose to use Logistic Regression, Decision Tree, and Neutral Network models to analyze our data set.

LOGISTIC REGRESSION	DECISION TREE	NEURAL NETWORK
A Logistic Regression Model is used to predict the likelihood of the categorical dependent variable using a independent variable. Logistic regression is used for solving categorical problems rather than regression problems.	A Decision Tree Model is an algorithm that classifies outcomes on a set of rules and conditions. This can be broken down to decision nodes and leaf nodes.	Neural Network Model is a machine learning tool used to mimic human like understanding and decision making through a connection of nodes.

REGRESSION ANALYSIS MODEL

We ran a logit regression analysis by keeping number of cancellations (is_cancelled) as the target variable and a few independent variables. Overall the model was significant as the p value is less than 0.05.

Most of the independent variables are insignificant as their p value is greater than 0.05 except lead time , deposit type and total_of_special_requests whose p value is less than 0.05.

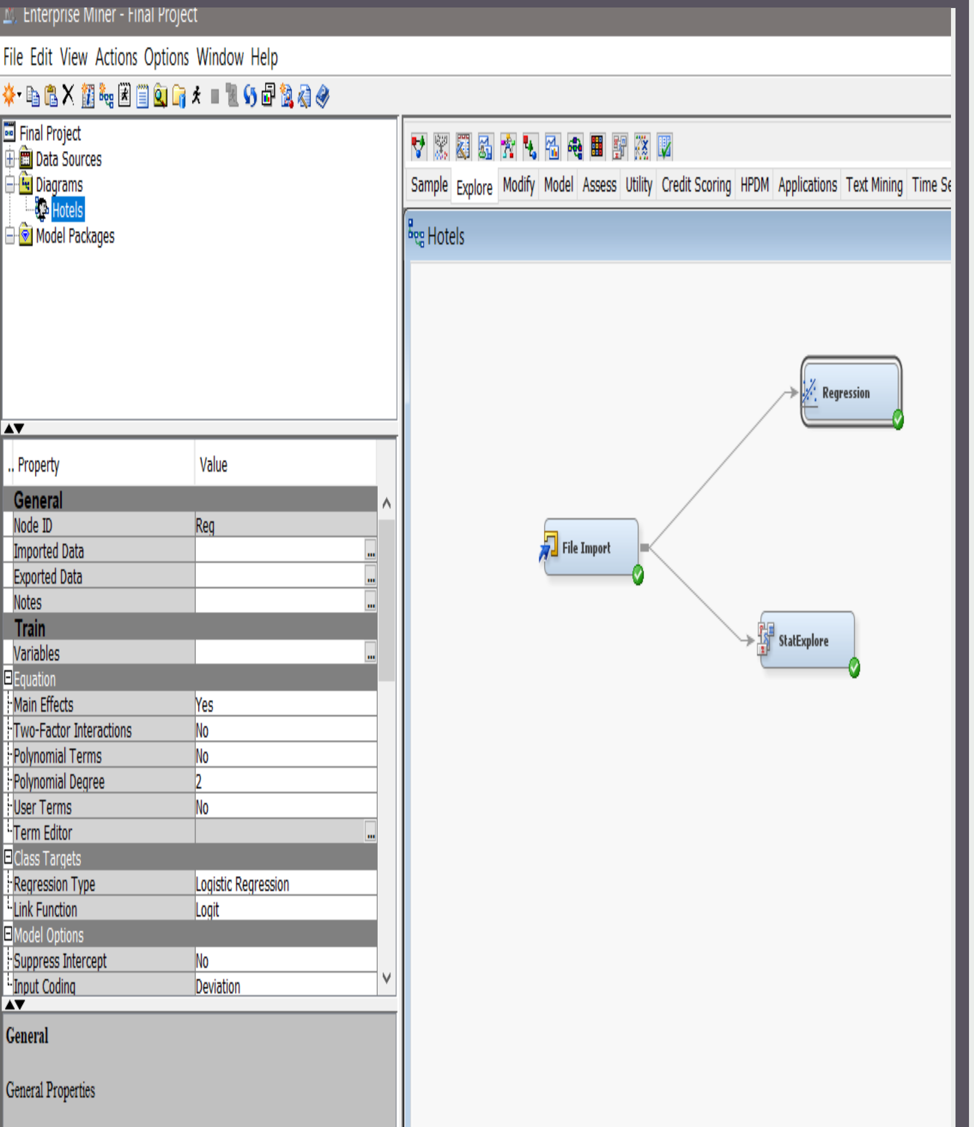
With an ASE of 0.14

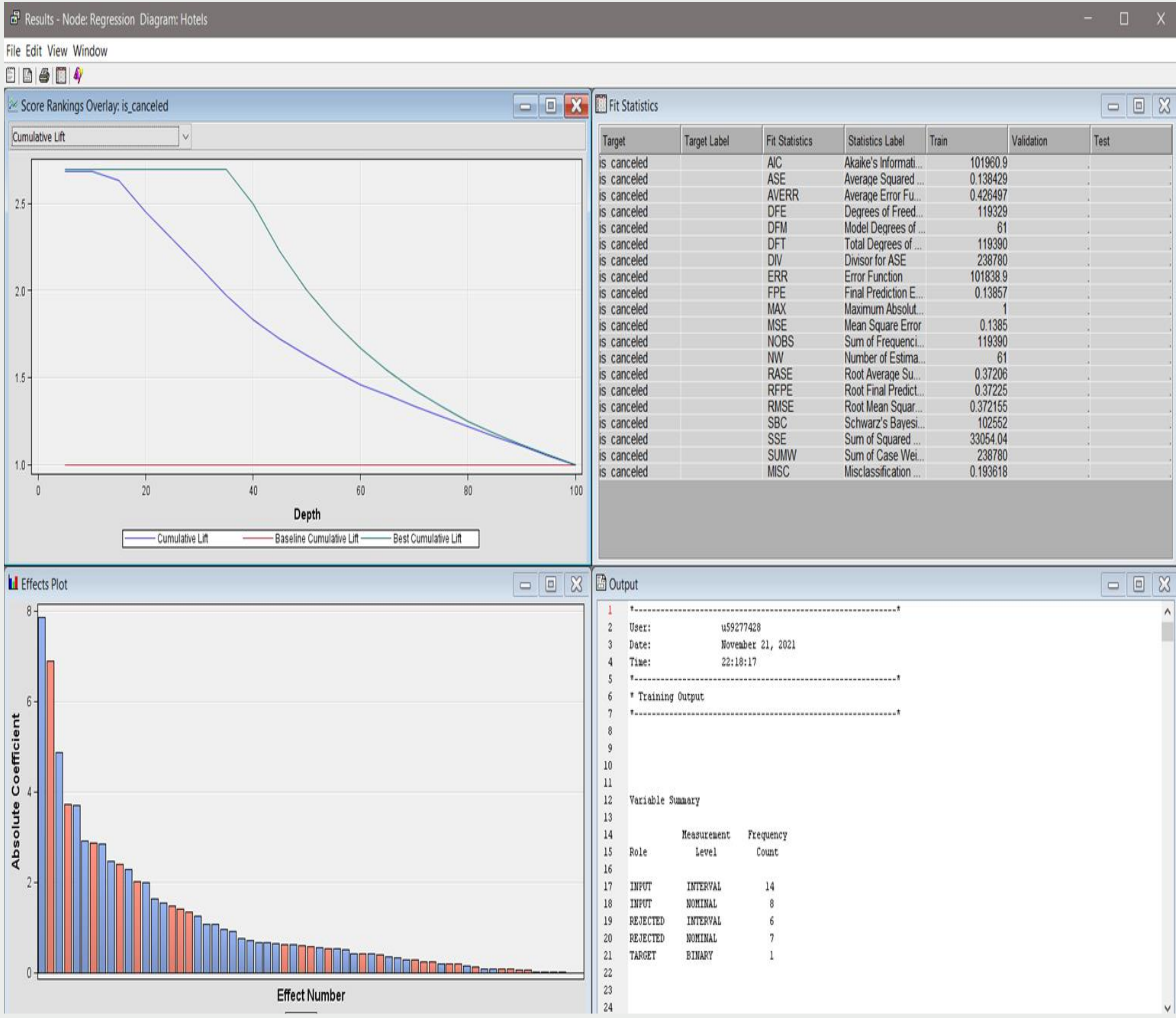
Fit Statistics

Target=is_canceled Target Label=' '

Fit Statistics	Statistics Label	Train
AIC	Akaike's Information Criterion	100437.54
ASE	Average Squared Error	0.14
AVERR	Average Error Function	0.42
DFE	Degrees of Freedom for Error	119328.00
DFM	Model Degrees of Freedom	62.00
DFT	Total Degrees of Freedom	119390.00

REGRESSION SEM





REGRESSION ANALYSIS

The analysis from our logistic regression model will provide hotel companies a better idea of the ‘**Probability of reservation cancellation’** under a few different scenarios as logit regression is based on probability or likelihood.

Hence, companies can adjust their reservation policy in terms of giving them lead time, keeping deposits etc for cancellations to better utilize their capacity and reduce profit lost

Output

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	1.6213	4.7217	0.12	0.7313		5.060
adr		1	0.00507	0.000239	450.04	<.0001	0.1412	1.005
adults		1	0.1409	0.0169	69.27	<.0001	0.0450	1.151
arrival_date_month	Apri	1	-0.1217	0.0678	3.22	0.0729		0.885
arrival_date_month	Augu	1	-0.0148	0.0444	0.11	0.7389		0.985
arrival_date_month	Dece	1	0.5546	0.1433	14.98	0.0001		1.741
arrival_date_month	Febr	1	-0.1040	0.1169	0.79	0.3736		0.901
arrival_date_month	Janu	1	-0.2881	0.1435	4.03	0.0446		0.750
arrival_date_month	July	1	-0.1659	0.0262	40.11	<.0001		0.847
arrival_date_month	June	1	-0.1272	0.0286	19.83	<.0001		0.881
arrival_date_month	Marc	1	-0.2943	0.0928	10.06	0.0015		0.745
arrival_date_month	May	1	-0.1346	0.0453	8.82	0.0030		0.874
arrival_date_month	Nove	1	0.3954	0.1184	11.15	0.0008		1.485
arrival_date_month	Octo	1	0.2711	0.0929	8.51	0.0035		1.311
arrival_date_week_number		1	-0.0150	0.00576	6.74	0.0094	-0.1122	0.985
assigned_room_type	A	1	-0.00848	4.8134	0.00	0.9986		0.992
assigned_room_type	B	1	-0.7011	4.8137	0.02	0.8842		0.496
assigned_room_type	C	1	-1.4569	4.8141	0.09	0.7622		0.233
assigned_room_type	D	1	-1.3594	4.8135	0.08	0.7776		0.257
assigned_room_type	E	1	-2.0860	4.8140	0.19	0.6648		0.124
assigned_room_type	F	1	-2.6942	4.8147	0.31	0.5758		0.068
assigned_room_type	G	1	-3.4664	4.8168	0.52	0.4717		0.031
assigned_room_type	H	1	-2.2879	4.8286	0.22	0.6356		0.101
assigned_room_type	I	1	-4.4488	4.8350	0.85	0.3575		0.012
assigned_room_type	K	1	-2.5386	4.8227	0.28	0.5986		0.079
assigned_room_type	L	1	7.1497	49.1973	0.02	0.8845		999.000
babies		1	0.2739	0.0855	10.27	0.0014	0.0147	1.315
booking_changes		1	-0.3660	0.0155	560.90	<.0001	-0.1316	0.694
children		1	0.2220	0.0250	79.06	<.0001	0.0488	1.249
customer_type	Contract	1	-0.2391	0.0572	17.46	<.0001		0.787
customer_type	Group	1	-0.4360	0.1258	12.01	0.0005		0.647
customer_type	Transient	1	0.5893	0.0450	171.13	<.0001		1.803
days_in_waiting_list		1	-0.00053	0.000483	1.19	0.2756	-0.00511	0.999
deposit_type	No Deposit	1	-1.8881	0.0815	537.07	<.0001		0.151
deposit_type	Non Refund	1	3.5624	0.1042	1167.85	<.0001		35.247
distribution_channel	Corporate	1	0.1740	0.0716	5.90	0.0151		1.190
distribution_channel	Direct	1	-0.4387	0.0761	33.19	<.0001		0.645
distribution_channel	GDS	1	-0.9348	0.1898	24.26	<.0001		0.393
distribution_channel	TA/TO	1	0.0625	.	.	.	.	1.065
hotel	City Hotel	1	-0.0833	0.00977	72.71	<.0001		0.920
is_repeated_guest		1	-0.6125	0.0843	52.85	<.0001	-0.0594	0.542
lead_time		1	0.00393	0.000099	1573.83	<.0001	0.2314	1.004
market_segment	Aviation	1	-0.7940	0.1849	18.44	<.0001		0.452
market_segment	Complementary	1	-0.0194	0.1462	0.02	0.8946		0.981
market_segment	Corporate	1	-0.8737	0.0831	110.58	<.0001		0.417
market_segment	Direct	1	-0.6417	0.0801	64.17	<.0001		0.526
market_segment	Groups	1	-0.7096	0.0394	323.76	<.0001		0.492
market_segment	Offline TA/TO	1	-1.3127	0.0269	2376.16	<.0001		0.269

DECISION TREE

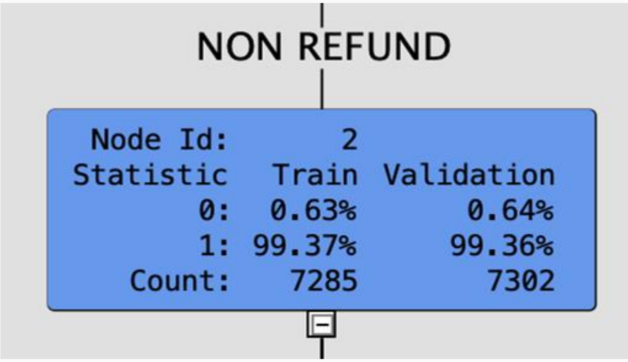
Data was partitioned using 50% Training, 50% Validation, 0% Test. The maximum branches were set to 3 and depth kept at 6. We attempted to increase the maximum branches but the decision tree became too hard to interpret.

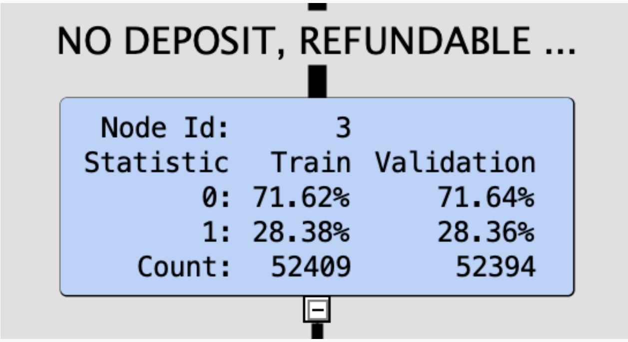
Rejected values included: agent, arrival date year, company, credit card, meal, name, phone number, previous cancellations, and reservation status due to missing values or irrelevance to the target variable.

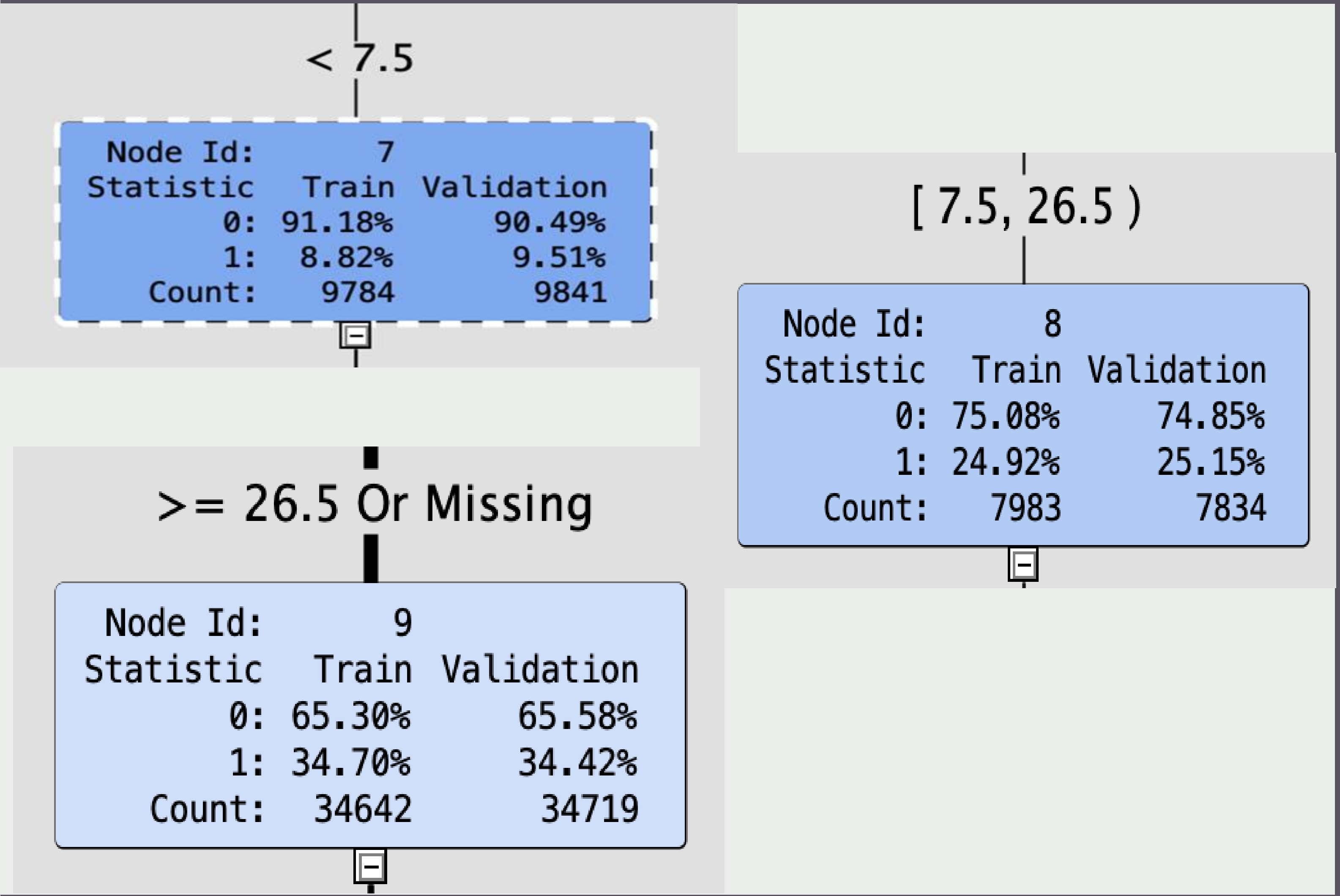
ASE = .128

DEPOSIT TYPE ANALYSIS

The first split is deposit type: non refund (node ID 2) or no deposit/refundable (node ID 3). From there, we can see that majority of those who had a non-refundable deposit were likely to cancel (99%) while the majority who had no deposit or a refundable one were less likely to cancel their reservation (~71%)





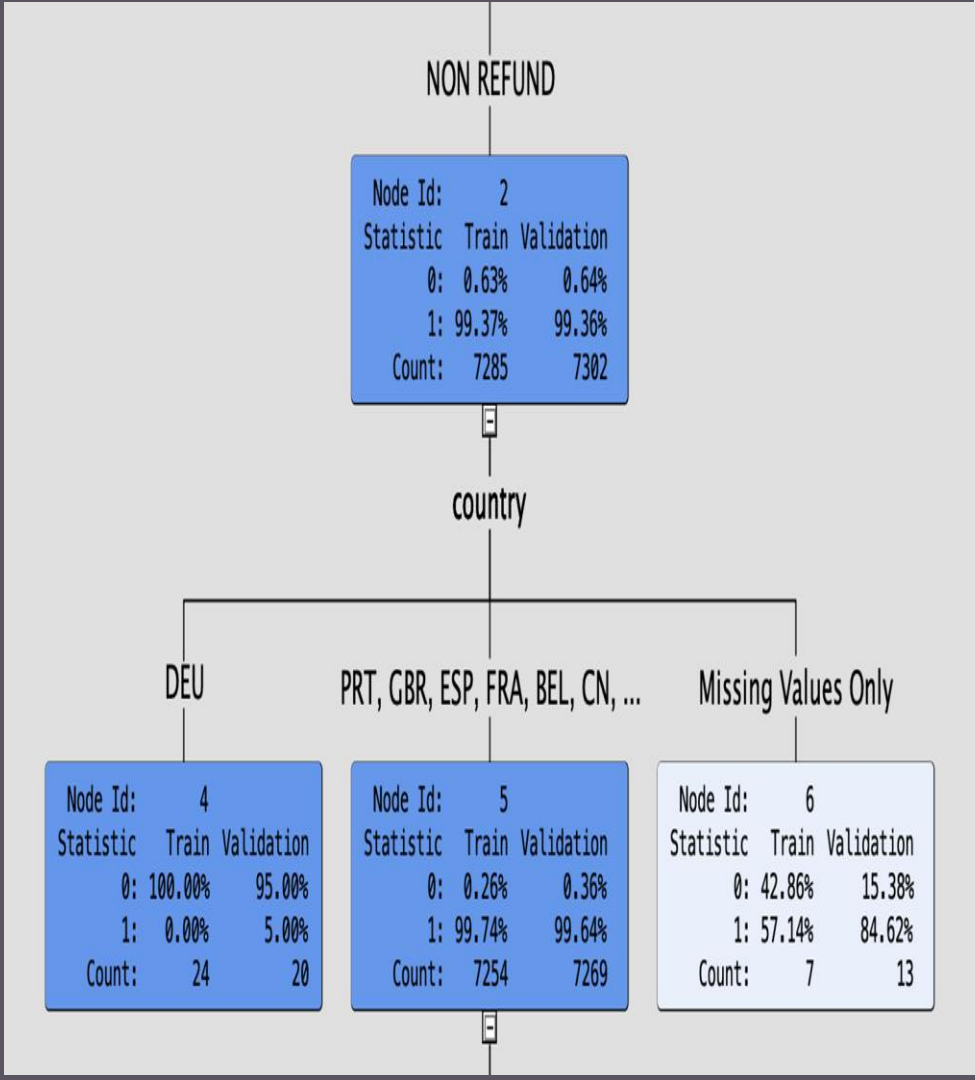


LEAD TIME ON CANCELLATIONS

The next split under Node ID 3 is lead time. This is broken up into <7.5 weeks, 7.5-26.5 weeks, and >= 26.5 weeks. From this we can interpret that with a shorter booking lead time, people are less likely to cancel their reservations (91% kept theirs) while those with longer lead times have a higher chance to cancel their bookings (~ 25% to 34% cancelled)

COUNTRY ON CANCELLATIONS

Under Node ID 2, we can see there is a split by country. We can see that those from DEU, are likely to keep their reservation (100% did) while those from Portugal, Great Britain, Spain, France, Belgium, and China are likely to cancel their reservation (over 99% did).



NEURAL NETWORK MODEL

The data was partitioned using a Training 60% , Validation 20% and Test 20% split

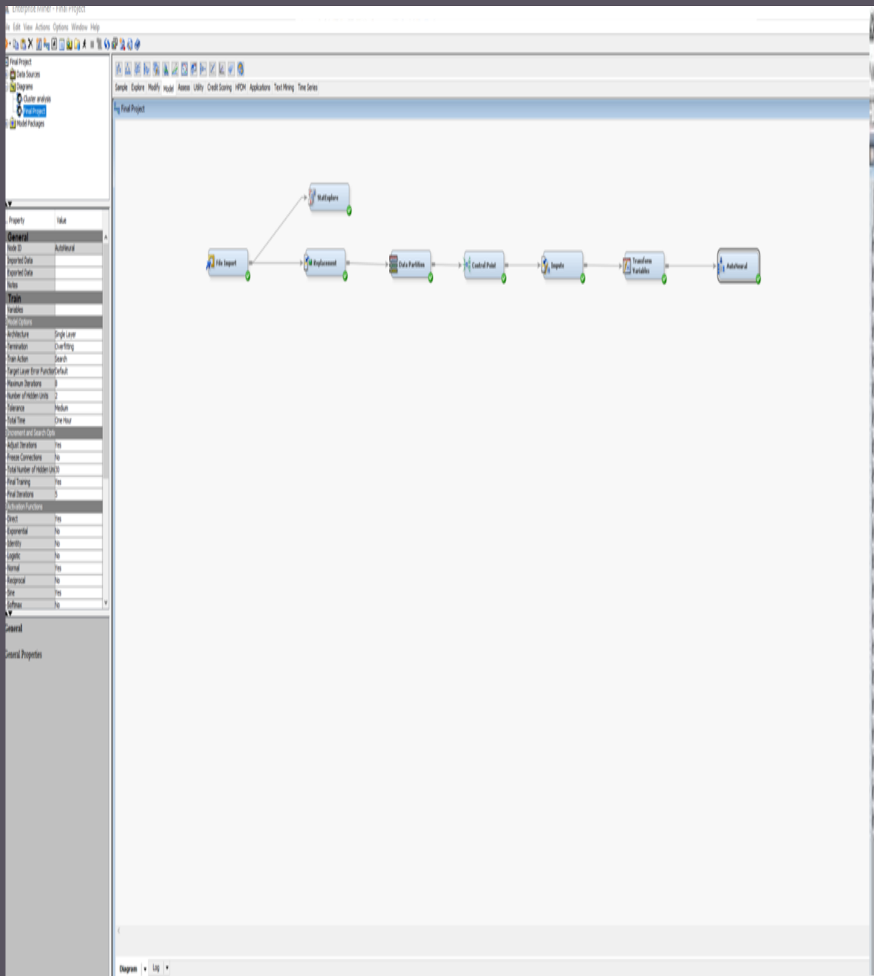
Our Neural Network Model used a tree surrogate imputation method for missing values

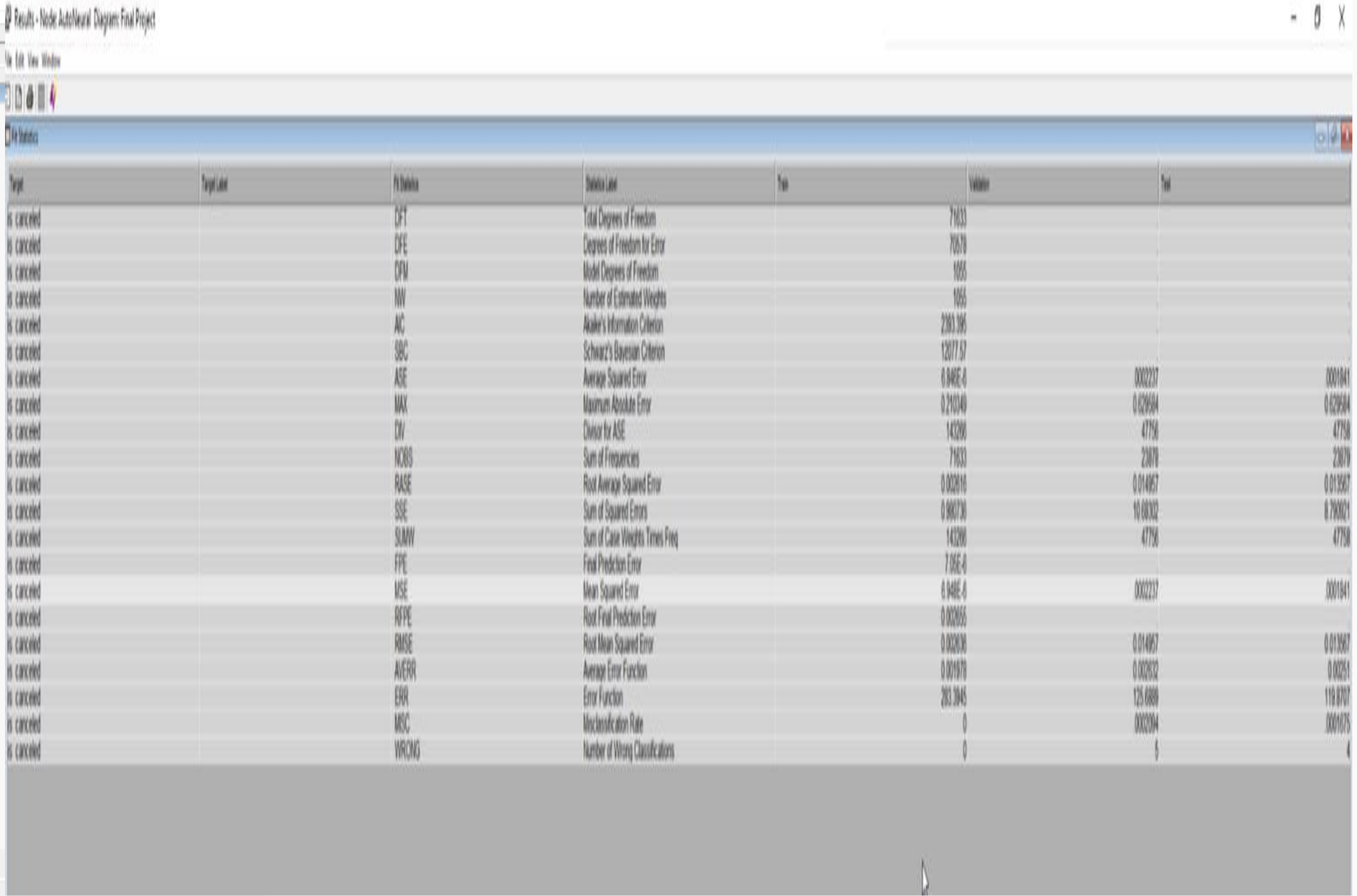
After transformation we did still see elevated skewness on a few variables

Using MSE we can see the values below after using 2 hidden units

- Train - .0000006958
- Validation - .0002237
- Test - .0001841

NEURAL NETWORK SEM





NEURAL NETWORK ANALYSIS

Our findings from the Neural Network found that the model produced a .0000168 ASE rate which tells us that the likelihood that a reservation is called is very low

In conclusion the model helps the hotel know that the likelihood of a reservation being cancelled is low and that they can operate and prepare for customers based on the number of reservations booked at the hotel

KEY FINDINGS

If we were using the average squared error to determine the best model of the three to show the likelihood of cancellation, the Neural Network Model would win by a landslide with a ASE of (.00001) compared to the logistical regression model and decision tree values of .14 and .128 respectively.

However, our group found the most detailed model to be the **Decision Tree.** Depending on the level of leaves and splits, the model was able to deduce the significance of each variable on our target, cancellation.

In regards to the Regression Model, dependent variables such as lead time and special requests showed significance with a p value of $<.05$.

MANAGERIAL/BUSINESS IMPLICATIONS

- Determine trends and reduce cancellations

For both regression analysis and the decision tree model, it was revealed that lead times played a significant factor in cancellations. Customers with shorter lead times are less likely to cancel. Therefore customers with longer lead times are likely to cancel. Interestingly, the percentage of customers that cancelled their reservation was significantly higher when a non refundable deposit was made.

- Catering to certain higher demographics and retain customer loyalty

Countries such as Germany were likely to keep their reservations (100% out of 24) While only 19 out of 7254 (0.26%) clients from a collective of countries, Portugal, Great Britain, Spain, France, Belgium, and China kept their reservations.

After analyzing these trends we recommend the city hotel and resort hotel to implement non refundable deposits for longer lead time reservations in order to recuperate costs spent on forecasting staff and preparations. We also recommend marketing towards clients from countries that are less likely to cancel. This can be done through advertisements or catering special requests to customs of their country. This is to build a loyal consumer base to retain customer loyalty to those who travel to these hotels.

CONCLUSION

The following models varied in results, however each played a significant role in our analysis. Logistic regression presented categorical dependent variables that made an impact on our target variable, cancellations. This revealed characteristics of variables that allowed us to target how we market and react to our consumer base.

The Decision Tree Model presented a view different leaves and decide the greatest manageable number. This allowed us to see the most significant splits that affected the number of cancellations. Not only so, it gave finite numbers and percentages which played a great impact on how we analyzed the previous consumer base of these hotels. We ultimately used this analysis to recommend business decisions to the hotels analyze.

Lastly, the neural network showed us the likeness of a cancellation to be low in instances. Which allowed us to believe normal conduction of business can be appropriate, however we used the other methods in order to not stay stagnant in business approaches, but to continually improve.

https://www.kaggle.com/mojtaba142/hotel-booking

https://www.sciencedirect.com/science/article/pii/S2352340918315191?via%3Dihub

REFERENCES