|  |
| --- |
| [Company name] |
| [Document title] |
| [Document subtitle] |

|  |
| --- |
| Submitted by:  Lorraine Ngamassi Deugoue – 301330685  Anum Haneef – 301111652  Sara Memon – 301420887  Zoha Awan - 301410599  [Date] |

INTRODUCTION:

The dataset we chose for our project aims to predict booking cancellations by analysing hotel reservations data. This dataset takes into account many factors some of which are : number of adults & children in the booking, the type of room and meal plan selected, the number of special requests made, how early in advance the booking was, how did they book the reservation and whether they were a returning customer or not etc.

ANALYSIS QUESTION:

How do we help hotel businesses predict which factor play a role in whether a booking is likely to be cancelled or not?

DATASET DESCRIPTION:

Our datset has been obtained from kaggle.com (cited at the end of the report). It has 17 variables including ID and Target Variable (Cancelled/Not Cancelled) and 36,286 observations.

DATA PREPARATION:

First, we look at our data set, the input variables, to avoid irrelevancy and redundancy. We made sure that the levels were related to the inputs and changed them when necessary. We decided to:

* Reject ‘P\_C’ and ‘P\_not\_C’ because they did not provide any relevant information and their values did not seem to explain anything.
* Although ‘date’ seemed important, it was impossible for us to change it in months for a better training, so it was rejected it as well.
* Then, we changed the ‘target’ level to ‘binary’.

The following screenshot details all the changes we made:

A screenshot of a computer

Description automatically generated

After importing our data set, we performed a replacement to assign the values 1 and 0 to our prediction target (1 for Canceled and 0 for Not\_Canceled because we want to predict the canceled booking for hotels):

A screenshot of a computer

Description automatically generated

Then, the next step was to partition the data, and we decided to split it 60:40:0 (Train, Validation, Test), as shown below:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

In fact, we decided to use more data to train the model than to validate it, because the model needs to see many examples of input data to learn how to make accurate predictions.

PREDICTIVE MODELING:

DECISION TREES:

To optimize our decision tree models, we ran 3 models: Maximal Tree (largest), Misclassification Tree (assessment, decision-misclassification), and ASE Tree (assessment, decision-average square error) :

A screenshot of a computer

Description automatically generated

MAXIMAL TREE:

Our largest (Maximal Tree) had the validation ASE of 0.116499 as shown below:

A screenshot of a computer

Description automatically generated

Then we ran the other two Trees (Misclassification & ASE) and compared the ASE as follows:

1. OPTIMAL DECISION TREE:

|  |  |  |
| --- | --- | --- |
| Name of Decision Tree | Average Squared Error | Assessment Type |
| Maximal Tree | 0.112838 | Largest |
| Misclassification Tree | 0.117816 | Misclassification |
| **ASE Tree** | **0.112791** | **Average Squared Error** |

Based on the above table, the validation ASE for ASE Tree appeared the lowest among the 3 models (**0.116219** versus 0.116499 for Maximal Tree and 0.119821 for Misclassification Tree). Therefore, **ASE Tree was assumed the optimal Decision Tree**.

A.1) ASE TREE:

The Complete Results of the ASE are explained bellow:

A screenshot of a computer

Description automatically generated

A diagram of a diagram

Description automatically generated with medium confidence

A diagram of a company

Description automatically generated

The screenshots of ASE Tree show us that following are the important variables for the model:

* lead time (0: 67.23 % and 1: 32.77%),
* special requests (0: 76.84 % and 1: 23.16%)
* and average price (0: 27.12 % and 1: 72.88%).

A white paper with black text

Description automatically generated

Moreover, the output table of ASE Tree showed us in variable importance, lead\_time (1.000 of validation importance), market\_segment\_type (0.6337 of validation importance), special\_requests (0.5313 of validation importance), and average\_price(0.4923 of validation importance) as important variables for the model.

DATA PREPARATION FOR REGRESSION:

For Regression based modeling we need to fix missing values, determine any outliers, normalise skewed data, and try to reduce the degrees of freedom to find out the optimal Regression model. All of these steps are shown below:

1. IMPUTATION

Even though we do not have missing values in our dataset, we decided to run ‘Impute’ as a precautionary measure to ensure there are no missing values that we may have missed. To add the Impute node we have ‘None’ to ‘Unique’ (type) and ‘Rejected’ to ‘Input’(role)

* 1. IMPUTATION RESULTS:

The stat explores of impute showed us that our data had some outliers and because of that some of our variables were skewed, such as: lead\_time (1.68), special\_requests (1.52), number\_of\_children (3.08).

**A screenshot of a computer

Description automatically generated**

1. CAP & FLOOR:

We then used cap and floor (standard deviation at 3) in an attempt to fix the skewed values:

A screenshot of a computer

Description automatically generated

**A screenshot of a computer

Description automatically generated**

As shown above, no changes were made in skewness after cap and floor.

1. TRANSFORM VARIABLES:

Then, we decided to use transform variable. For the three variables (lead\_time, special\_requests & number\_of\_children) to be fixed we had to carry out all the transformation methods, such as log (which made no difference) and then log10 (which fixed some but not all) as shown below:

1. Results after Log10 transformation:

**A screenshot of a computer

Description automatically generated**

We can see from the results that some of the skewness decreased with Log10. For example, lead\_time (1.58 ) became -0.67, special\_requests (1.52) became 1.19, however number\_of\_children (remained skewed with a value of 3).

Therefore, we continued to transform variables using the best method:

As you can see from the results, the skewness disappeared from all the variables, however the three skewed variables were optimized into four bins.

1. RECODE DUMMIES:

To reduce the degrees of freedom (dimensions), we needed to examine the categories that were like each other in order to reduce the redundancy of the data. For that, we determined that:

* Room\_Type 1 & 3 should be merged into a new category Room\_Type 13 as Room\_Type 3 had only 5 observations and Room\_Type 1 had the largest number of observations (16866). There was no explanation as to what was different about the room types therefore, we merged these two to ensure that our data average is least affected by this recoding.
* We used the same logic for Meal Plan 1 & 3, as Meal Plan 3 had only 2 observations. Therefore, we merged it with the largest observations: Meal Plan 1 and made Meal Plan 13 as the new category for both. Again, there was no explanation for the different Meal Plans, so we decided to do what least impacted the average:

A screenshot of a computer

Description automatically generated

REGRESSION MODELS:

We ran four type of Regression models after recoding our categories: full, forward, backward, and stepwise regression.

FULL REGRESSION MODEL:

A screenshot of a computer

Description automatically generated

The ASE of Full Regression was 0.135473. After that we decided to run the three Regression Models based on the Model Selection Criterion: (Forward, Backward and Stepwise).

OPTIMAL REGRESSION MODEL:

The resulting ASEs for all the Regression models were as follows:

|  |  |  |
| --- | --- | --- |
| Name of Regression Model | Average Squared Error | Model Selection |
| Full Regression | 0.135473 | None - Full |
| **Forward Regression** | **0.135201** | **Forward** |
| **Backward Regression** | **0.135383** | **Backward** |
| **Stepwise Regression** | **0.135274** | **Stepwise** |

The results of the regression models were similar except for Full Regression which had slightly higher ASE than the others.

Since Forward Regression, Backward Regression and Stepwise Regression had similar values of ASE, we chose **Stepwise Regression** as the best and interpreted its Odds Ratios.

ODDS RATIO INTERPRETATION FOR STEPWISE REGRESSION MODEL:

. A close-up of a computer code

Description automatically generated

The interpretation of the Odds Ratios of the Stepwise Regression Output is discussed below:

* The bookings with lead time of 9.5 and below are to be 98.6% less likely to cancel than the bookings with lead time of 151.5 and high.
* The bookings with lead time of 9.5 to 100.5 are 94.4% less likely to cancel than the bookings with lead time of 151.5 and high.
* The bookings for lead time of 100.5 to 151.5 are 87% less likely to cancel than the ones who have the lead time of 151.5 and high.
* The bookings with 1 child are 40.8% less likely to cancel than the bookings who have more than 1 children.
* The bookings who have 0 special request are 999 times more likely to cancel than the bookings who have 2 or more special requests.
* The bookings who have 0 to 1 special requests are 999 times more likely to cancel than the bookings who have more than 2 special requests.
* The bookings who have 1 to 2 special requests are 999 times more likely to cancel the bookings who have more than 2 special requests.
* The bookings who have more average price are 1.8% more likely to cancel than the ones who have less average price.
* The bookings that have more numbers of adults are 20.8% more likely to cancel than the ones who have a smaller number of adults.
* The bookings that have more number of weekends are 10.1% more likely to cancel than the ones who have less number of weekends.
* The bookings with room types 1 and 3 are 3.677 time more likely to cancel than room type 7.
* The bookings with room types 2 are 2.454 times more likely to cancel than room type 7.
* The bookings with type 4 are 42.8% more likely to cancel than room type 7.
* The bookings with room type 6 are 29% more likely to cancel than room type 7.
* Bookings with mealtimes 1 or 3 are 29.2% less likely to cancel than the people who do not have their meal selected.
* The booking that selected meal plan 2 are 6% more likely to cancel than the ones who have not selected their meals.
* The one who do not have a parking spot is 4.789 times more likely to cancel than the one who have a parking spot.
* The bookings that were made in market segment type aviation are 70.7% more likely to cancel than the ones who booked online.
* The bookings that were made in market segment type complementary are 99.9% or even more likely to cancel than the ones who booked online.
* The bookings that were made in market segment type corporate are 57.7% less likely to cancel than the ones who booked online.
* The bookings that were made in market segment type offline are 84.1% less likely to cancel than the ones who booked online.
* The guest who are coming for the first time are 6.781 times more likely to cancel than the ones who have the booking in the past.

NEURAL NETWORK:

After Regression we modeled based on Neural Network. First, we attached the NN Node to Impute, which gave us an NN Impute Model. After that, we ran NN Cap & Floor, NN Transform and NN Recode (by connecting to Cap & Floor, Transform and Recode Dummies nodes). The resulting ASEs were as follows:

|  |  |  |
| --- | --- | --- |
| Name of Neural Network | Average Squared Error | No. of Iterations |
| NN Impute | 0.129721 | 67 |
| NN Cap & Floor | 0.130146 | 64 |
| **NN Transform** | **0.119829** | **52** |
| NN Recode | 0.120561 | 97 |

Based on this table, our best Full Neural Network model appeared to be NN Transform with the least ASE.

FULL NEURAL NETWORK:

A screenshot of a computer

Description automatically generated

REGRESSION NEURAL NETWORK:

For Regression Neural Network, we ran four models with Stepwise Regression, as the ASE for all three Regression models was the same. We started with 3 hidden variables and ran the Neural Network node till 6 hidden variables (6H). After that, the ASE started increasing so we stopped there. Our best Regression Neural Regression Network with lowest ASE appeared to be with 6 hidden variables. The comparisons of ASEs are as follows:

|  |  |  |
| --- | --- | --- |
| Name of Regression Neural Network | Average Squared Error | No. of Iterations |
| NN Forward 3H | 0.121713 | 61 |
| NN Forward 4H | 0.119978 | 72 |
| **NN Forward 5H** | **0.116519** | **62** |
| NN Forward 6H | 0.118593 | 98 |

A screenshot of a computer

Description automatically generated

Our optimal Neural Network model out of all the NN nodes appeared to be NN Forward 5H – with 5 hidden variables, based on ASE.

MODEL ASSESSMENT:

In the final step, we performed model assessment to find out which model is the best based on ROC.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

By connecting all the nodes to the model comparison, we can see that NN forward 5H has the highest ROC index (0.897). However, based on the analysis ASE tree would be the best model as it has the lowest Average square error (0.112791).

SUMMARY:

To summarize, our best model appeared to

CONCLUSION & RECOMMENDATIONS:

In conclusion we can say that the two models with highest Validation ROC Index are ASE TREE and

Based on the ASE Tree Analysis, following insights can be drawn:

* The highest chance of cancellation is with bookings who with between 151.5/Missing to 6.5/Missing days of lead time with special emphasis on people who made 0.5-1.5 special requests, booked online and did not need a car parking space with a validation count of 3878 observations.
* The next best category appears to be for bookings that had between 151.5-90.5/Missing days of lead time, more than 1.5 special requests and less than 3.5 number of weeknights. The validation count of this category is 1860 observations.
* The third important category is bookings with less than 0.5 special requests with more or equal to 9.5 days of lead time, booked online, with avg. price less than or equal to 109.29 and did not need a parking space. This has 1623 observations.

From the business perspective, we can say that:

* Number of days leading up to the actual booking, how the booking is made: Online / Offline / Corporate, number of special requests, avg. price of the booking play significant role in the prediction of booking cancellations.
* The more the special requests they have, the less the leading time for the booking, and less the avg. price , the more the chances of booking not being cancelled.

CITATION:

*Hotel booking cancellation prediction*. (2023, December 3). Kaggle. <https://www.kaggle.com/datasets/youssefaboelwafa/hotel-booking-cancellation-prediction>

APPENDIX:

Dataset Approval Email:

A screenshot of a phone

Description automatically generated