**DATA MINING AND PREDICTIVE ANALYTICS**

**PROJECT REPORT ON HOTEL BOOKING DATASET**



**CONTENTS**

1. **Introduction**
   1. Background
   2. Business Question
   3. Plan for answering question
   4. Objective
2. **Data collection process**
   1. Variable Explanation
   2. Data Type of variables, possible range of values and relevance to business question
3. **Data summarization**
4. **Data visualization**
   1. Histogram of Lead Time
   2. Bar chart for room\_type
   3. Box plot for stays\_in\_week\_nights and is\_cancelled
   4. Scatter plot for previous\_cancellations and is\_repeated\_guest
5. **Data Analysis**

5.1 Technique 1: Logistic Regression

5.2 Technique 2: KNN

1. **Implementation**
   1. Technique 1: Logistic Regression
      1. Implementation
      2. Result and Analysis
      3. Interpretation of Validation dataset
      4. Classification of booking cancellation for a new customer
   2. Technique 2: KNN
      1. Implementation
      2. Result and Analysis
      3. Interpretation of Validation dataset
      4. Classification of booking cancellation for a new customer
2. **Result and Analysis**
3. **Comparison**
4. **Summary and Recommendations**

**DELIVERABLE – 3**

**HOTEL BOOKING DATASET**

**INTRODUCTION:**

In the dynamic landscape of the hospitality industry, understanding and predicting the customer satisfaction plays an important role for effective management and resource allocation. In this project, we aim to explore the predictive power of various features in hotel booking dataset to determine whether the booking is likely to be cancelled or not.

**BACKGROUND:**

The hospitality sector faces challenges in optimizing the room occupancy and managing the resources efficiently. One of the critical aspects influencing this balance is the phenomenon of booking cancellations. Understanding the factors that contribute to cancellations and developing the prediction and classification models can increase the hotel’s ability in managing resources, allocating resources efficiently and improves the customer satisfaction.

**BUSINESS QUESTION:**

Can we predict whether a booking will be cancelled or not based on features like lead time, number of special requests, room type and previous cancellations and many other columns.

To answer this business question, we are using two data mining techniques: Logistic Regression and K-Nearest Neighbors (KNN). These techniques allow us to analyze patterns within the dataset and build models capable of predicting the likelihood of booking cancellations.

**PLAN FOR ANSWERING QUESTION:** Based on the business question, we are predicting whether a booking will be cancelled or not. Output of the business question will be 0 or 1(yes/no) which is in binary format, so we are using logistic regression and KNN in order to predict the output. We will compare both the outputs of logistic regression, KNN and finds the best method of two.

**OBJECTIVE:**

* Apply Logistic Regression and K-Nearest Neighbors to predict hotel booking cancellations.
* Analyzing and comparing the results obtained from these two data mining techniques.
* Providing insights and recommendations based on the predictive models.

**DATA COLLECTION PROCESS:**

We have selected the Hotel Booking Dataset in order to analyze the data for performing different data mining techniques.

Dataset source: We have taken Hotel Booking Dataset from Kaggle. Below is the link given for Hotel Booking Dataset.

<https://www.kaggle.com/datasets/saadharoon27/hotel-booking-dataset>

Hotel Booking dataset structured with 22 columns, each representing the distinct aspect of hotel reservations. It contains information about various aspects of hotel booking including arrival details, customer demographics, booking changes, previous cancellations and ultimate cancellation status.

We have converted hotel,arrival\_date\_month,reserved\_room\_type and assigned\_room\_type column values into number format in order to increase the accuracy.

**Variable Explanation:**

* **hotel**: The type of hotel, either "City Hotel" or "Resort Hotel". We have converted the hotel column values into binary format.

**City Hotel – 0**

**Resort Hotel – 1**

* **lead\_time**: Number of days between booking and arrival.
* **arrival\_date\_year**: Year of arrival date.
* **arrival\_date\_month**: Month of arrival date. We have converted the arrival\_date\_month column values into numerical format (**January -1, February -2, March – 3….)**
* **arrival\_date\_week\_number**: Week number of arrival date.
* **arrival\_date\_day\_of\_month**: Day of the month of arrival date.
* **stays\_in\_weekend\_nights**: Number of weekend nights (Saturday or Sunday) the guest stays.
* **stays\_in\_week\_nights**: Number of weekday nights (Monday to Friday) the guest stays.
* **adults**: Number of adults.
* **children**: Number of children.
* **babies**: Number of babies.
* **is\_repeated\_guest**: Binary value indicating whether the guest is a repeated guest (1) or not (0).
* **previous\_cancellations**: Number of previous booking cancellations.
* **previous\_bookings\_not\_canceled**: Number of previous bookings not cancelled.
* **reserved\_room\_type**: Code of room type reserved. We have converted the reserved\_room\_type column values into number format (A-1,B-2,C-3).
* **assigned\_room\_type**: Code of room type assigned at check-in. We have converted the assigned\_room\_type column values into number format (A-1,B-2,C-3).
* **booking\_changes**: Number of changes/amendments made to the booking.
* **days\_in\_waiting\_list**: Number of days in the waiting list before booking.
* **required\_car\_parking\_spaces**: Number of car parking spaces required.
* **total\_of\_special\_requests**: Number of special requests made.
* **name**: Guest's name.
* **phone-number**: Guest's phone number.
* **is\_cancelled**: whether the booking is cancelled or not.

**Data Type of variables, possible range of values and relevance to business question:**

* **hotel** 🡪 It is numerical data and possible range of values are 0 and 1.
* **lead\_time** 🡪 It contains the numerical data which is used to predict whether the booking cancelled or not.
* **arrival\_date\_year** 🡪 **This column contains numerical data which shows the information about year.**
* **arrival\_date\_month** 🡪 **It is in numerical format and possible range of values are 1,2,3…. If the month of booking is during holiday time, then the probability of booking getting cancelled is less.**
* **arrival\_date\_week\_number** 🡪 **It shows the week number in the numerical format.**
* **arrival\_date\_day\_of\_month** 🡪 **It contains the numerical data which shows the day of month of arrival date.**
* **stays\_in\_weekend\_nights** 🡪 **It contains the numerical data which shows the number of weekend nights the guest is going to stay in hotel.**

**If the guests are going to stay on weekend nights, the chances of booking cancelled will be less.**

* **stays\_in\_week\_nights** 🡪 **It is in numerical form which shows the number of weekday nights guest stays in hotel. During weekdays, the probability of booking cancelled is more.**
* **Adults** 🡪 It is in numerical format which shows the number of adults.
* **Children** 🡪 It is in numerical format which shows the number of children.
* **Babies**🡪 It is in numerical format which shows the number of babies.
* **is\_repeated\_guest** 🡪 It contains the numerical data and if the guest comes repeatedly to hotel, the probability of cancellation is less.
* **previous\_cancellations** 🡪 **It contains the data in number format, and we can predict the cancellation based on previous cancellations data. If number of previous cancellations is more than chance of cancellation is also more.**
* **previous\_bookings\_not\_canceled** 🡪 **It contains the data in number format, and we can predict the cancellation based on previous bookings not cancelled data. If number of previous bookings not cancelled is more, then chance of cancellation is less.**
* **reserved\_room\_type** 🡪 **This column contains the data in number format. Guests can reserve the room type if they need.**
* **assigned\_room\_type** 🡪 **This column contains the data in number format. Hotel management can assign the room to the guests as per the availability. If the reserved\_room\_type and assigned\_room\_type is different, then the probability of cancellation is more.**
* **booking\_changes** 🡪 **This column contains the data in number format which shows the number of changes made to the booking. If number of booking\_changes is more, then the probability of cancellation is more.**
* **days\_in\_waiting\_list** 🡪 It contains the numerical data which shows the number of days in waiting list. If number of days in waiting is more, then the chance of cancellation is more.
* **required\_car\_parking\_spaces** 🡪 **It contains the numerical data which shows the required car parking spaces. If hotel management delays to provide the number of parking spaces asked by the guest, then cancellation chances are more.**
* **total\_of\_special\_requests** 🡪 It contains the numerical data showing the number of special requests made by the guests. If hotel management are not able to fulfill the special requests, then booking cancellations is more.
* **name** 🡪 This column contains the guest’s name in the text format.
* **phone-number 🡪 It contains the data in number format which shows the phone number of the guest.**

**Target Variable: “is\_cancelled”** column is the target variable which indicates whether the booking is cancelled or not. This binary variable serves as the focal point for enabling the classification of bookings into two categories cancelled (1) and not cancelled (0).

**Dataset Splitting:** This dataset was split into training and testing set. The training set was used to build and train the predictive model, while the testing set served as the independent dataset to evaluate the model’s performance.

**DATA SUMMARIZARION:**

* We have performed descriptive statistics on the hotel\_booking dataset in order to provide an overview of the data’s key characteristics.
* Descriptive statistics include mean, standard error, median, mode, standard deviation, sample variance, kurtosis, skewness, range, minimum, maximum, sum and count of the variables.
* We have two categorical variable and 20 numerical variables which includes phone number of the customers.
* Categorical variable is name of the customer.
* We have calculated the descriptive statistics of all the numerical variables below.

Table 1: Descriptive Statistics of hotel and lead\_time variables:



Table 2: Descriptive Statistics of arrival\_date\_year and arrival\_date\_month variables:



Table 3: Descriptive Statistics of arrival\_date\_week\_number and arrival\_date\_day\_of\_month variables:



Table 4: Descriptive Statistics of stays\_in\_weekend\_nights and stays\_in\_week\_night variables:



Table 5: Descriptive Statistics of adults and children variables:



Table 6: Descriptive Statistics of babies and is\_repeated\_guest variables:



Table 7: Descriptive Statistics of previous\_cancellations and reserved\_room\_type variables:



Table 8: Descriptive Statistics of assigned\_room\_type and booking\_changes variables:



Table 9: Descriptive Statistics of days\_in\_waiting\_list and required\_car\_parking\_spaces variables:



Table 10: Descriptive Statistics of total\_of\_special\_requests and is\_canceled variables:



**Overview of data’s key characteristics:**

* Variables with a large standard deviation or a wide range between the minimum and maximum values may be considered as the key variables in the data set. They have a significant spread of data points, which can provide valuable information.
* According to the descriptive statistics of numerical variables performed above, we can say that lead\_time, arrival\_date\_week\_number, arrival\_date\_day\_of\_month, stays\_in\_weekend\_nights, stays\_in\_week\_nights, previous\_cancellations, reserved\_room\_type, assigned\_room\_type, booking\_changes, days\_in\_waiting\_list, total\_of\_special\_requests have wide range of difference between maximum and minimum values in the hotel\_booking dataset. So, these variables are considered as the data’s key variables.
* According to the business question, variables like lead\_time, total\_of\_special\_requests, assigned\_room\_type, reserved\_room\_type, previous\_cancellations, is\_cancelled, is\_repeated\_guest and days\_in\_a\_waiting\_list are likely to be more crucial in this dataset.
* We need to visualize the data in order to explore the relationships between the variables.

**DATA VISUALIZATION:**

**Histogram of Lead\_time:**

* Hotel variable converted to text format in order to under the visualizations correctly. If Hotel\_type = 1 🡪 Resort and Hotel\_type = 0 🡪 City.
* Visualization: A histogram showing the distribution of lead time for both cancelled and non – cancelled flight bookings.
* We can understand that if there is any pattern in lead times that correlates with cancellations.

A screenshot of a computer

Description automatically generated

**Figure 1: Histogram**

From the histogram, we can conclude that lead\_time(number of days between booking and arrival) is more for resort type hotel. So, we can predict that number of cancellations for resort type hotel is more when compared to city type hotel.

**Bar chart for room\_type:**

* We have two types of room in the dataset like assigned\_room\_type and reserved\_room\_type.
* If the assigned\_room\_type and reserved\_room\_type are different, then we can predict the cancellation if the customer is not satisfied with the assigned\_room\_type.
* Assigned\_room\_type variable is more important here in order to predict the cancellations.
* Visualization: Bar chart showing the count of booking for each room\_type separated by the cancellations.

A screenshot of a computer screen

Description automatically generated

Figure 2: Barchart

* If is\_cancelled = 0, then the booking is not cancelled and is\_cancelled = 1 then the booking is cancelled.
* From above bar chart, we can predict that number of cancellations for resort type hotel is less when compared with the non-cancelled bookings.
* We can predict that the probability of booking getting cancelled for resort type hotel is less.
* For city type hotel, the number of booking cancellations and non-cancellations both are approximately same. So, the probability of getting cancelled is ½ and the probability of not getting cancelled is also ½.

**Box plot for stays\_in\_week\_nights and is\_cancelled:**

* Visualization: Box plot showing the distribution of week\_night stays of the customers for both cancelled and non-cancelled bookings.
* This visualization helps us to identify whether the week\_night stays affects the likelihood of cancellations.

A screenshot of a graph

Description automatically generated

Figure 3: Boxplot

**For Hotel\_tye = City:**

* We can see that the notches of the box plot are not overlapping. So, median will be different for them.
* We can observe that the distribution of data points is more for cancelled bookings. So, we can predict that if the customer booked for stay during week nights then the chance of booking gets cancelled is more.
* In this scenario, Booking cancellations and stay\_in\_week\_nights are directly proportional to each other.

**For Hotel\_type = Resort:**

* We can say that notched of the box plot are not overlapping with each other. So, median values will be different.
* It seems to be data distribution is almost same for cancelled and non-cancelled bookings but the maximum value is very high for cancelled bookings which means upper whisker is more.
* We can predict that if the customer booked for stay during week nights then the chances of booking bets cancelled is more.

We can conclude that for both hotel\_type = Resort and City, if the customer booked for stay in weeknights, the chance for booking gets cancelled is more.

**Scatter plot for previous\_cancellations and is\_repeated\_guest:**

* Visualization: A scatter plot with previous\_cancellations on y-axis and is\_repeated\_guest on x-axis shows whether there is any relationship between them.
* From the below scatter plot, we can say that there is a no relationship between the previous\_cancellations and is\_repeated\_guest variables in the dataset.
* That means repeated guests will not cancel the bookings more. There might be n number reasons for not cancelling booking. Some of them are mentioned below.
* Positive past experiences
* Loyalty program and rewards
* Familiarity and comfort
* Special requests and preferences
* Transparent and fair policies
* Quality of service
* Effective communication between customer and hotel management

A screen shot of a computer

Description automatically generated

Figure 4: Scatterplot

**DATA ANALYSIS:**

We are using two data mining techniques such as logistic regression and KNN in order to predict the outcome for hotel booking cancellations.

**TECHNIQUE 1 – LOGISTIC REGRESSION:**

* Logistic Regression is widely used statistical method for predicting the binary outcomes. Context of our business question is predicting whether booking will be cancelled or not – Logistic regression is suitable because it models the probability of binary outcome.
* The logistic function employed in this technique ensures that the predicted values will be binary (0 or 1) making it particularly appropriate for classification problems.

**Why Logistic Regression:**

* **Interpretability:** Logistic Regression provides clear and interpretable results. The coefficients of the model allow us to understand the impact of each feature on the likelihood of booking being cancelled.
* **Assumption of Linearity:** Logistic Regression assumes a linear relationship between the independent variables and log-odds of dependent variable. This assumption aligns well with our exploratory understanding of how certain features might influence booking cancellations.
* **Probability Estimation:** Logistic Regression outputs probabilities, which can be valuable for decision making. By setting a threshold, we can convert these probabilities into binary predictions.

**TECHNIQUE 2: K-NEAREST NEIGHBORS(KNN):**

* KNN is non parametric and instance-based learning algorithm used for both classification and regression tasks.
* The “K” in KNN represents the number of nearest neighbors to consider when making a prediction.

**Why KNN:**

* **Non-Linearity:** KNN doesn’t assume any underlying structure in the data. This flexibility allows it to capture non-linear relationship between feature and target variable which might be present in the hotel booking dataset.
* **Local Patterns:** KNN makes predictions based on the majority class of nearby instances. This can be advantage in scenarios where certain local patterns or clusters influence the likelihood of booking being cancelled.
* **No Assumptions about Data Distribution:** KNN makes minimal assumptions about the data distribution, making it robust in situations where the underlying data structure is not well defined.

**IMPLEMENTATION:**

**TECHNIQUE 1 - LOGISTIC REGRESSION:**

* Logistic regression is used for binary classification problems. The output is transformed using the logistic function which ensures the output is 0 or 1. The target variable is categorical and binary.
* **Data Splitting:** The dataset is divided into training and validation set used for model training and a validation set used for model evaluation. The split ratio is 60 – 40 means we have divided 60 percentage of data into training set and 40 percentage of data into validation set.
* In this case, we have partitioned the data into 60 percent into training data and 40 percent data into validation dataset. We have to rescale the data with standardization.
* **Feature Selection:** We have considered all the below mentioned features in order to perform the logistic regression except arrival\_date\_year because the value for that variable is same for all records so, there will not be any effect.

1. Lead\_time
2. Arrival\_date\_month
3. Arrival\_date\_week\_number
4. Arrival\_date\_day\_of\_month
5. Stays\_in\_weekend\_nights
6. Stays\_in\_week\_nights
7. Adults
8. Children
9. Babies
10. Is\_repeated
11. Previous\_cancellations
12. Booking\_changes
13. Days\_in\_waiting\_list
14. Required\_car\_parking\_spaces
15. Total\_special\_requests
16. Is\_cancelled

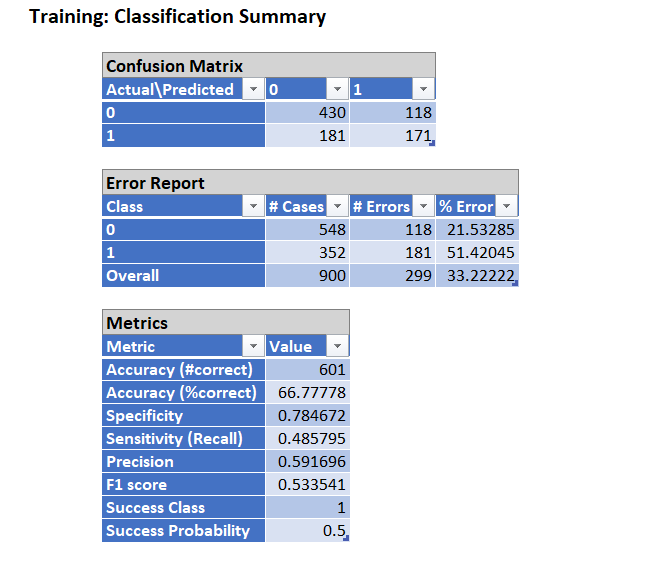
* **Data Preprocessing**: All the above features are considered and for two features like reserved\_room\_type and assigned\_room\_type, we have created the dummy variable.

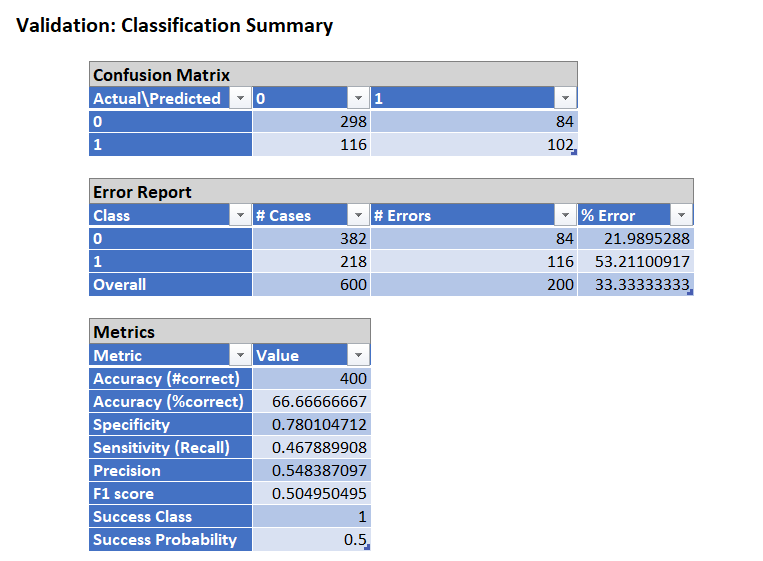
For reserved\_room\_type, we have three categories like 1, 2, 3. So, we require n-1 dummy variables means 2 dummy variables are created as reserved\_room\_typeD1 and reserved\_room\_typeD2.

For assigned\_room\_type, we have three categories like 1, 2, 3. So, we require n-1 dummy variables means 2 dummy variables are created as assigned\_room\_typeD1 and assigned\_room\_typeD2.

* **Model Training:** The logistic regression model is trained on the training set using the selected features. The model learns the relationship between the features and the target variable ('is\_canceled').
* **Model Evaluation:** The trained model is then evaluated on the validation set using performance metrics such as accuracy, precision, recall, F1 score.
* Cut off probability is set to 0.5. We have attached the confusion matrix and the classification summary for both training dataset and validation dataset below.

**RESULT AND ANALYSIS:**





**INTERPRETATION FOR VALIDATION DATASET:** A confusion matrix is a table that summarizes the performance of a classification algorithm. In this case, we have a binary classification problem with two classes: 0 (not canceled) and 1 (canceled).

**True Positives (TP):** 102 bookings were correctly predicted as canceled.

**True Negatives (TN):** 298 bookings were correctly predicted as not canceled.

**False Positives (FP):** 84 bookings were incorrectly predicted as canceled.

**False Negatives (FN):** 116 bookings were incorrectly predicted as not canceled.

* **Accuracy**: The model correctly predicted 66.67% of all cases.
* **Specificity**: The ability of the model to correctly identify non-canceled bookings is 78.01%.
* **Sensitivity (Recall):** The ability of the model to correctly identify canceled bookings is 46.79%.
* **Precision:** Among the predicted canceled bookings, 54.84% were actually canceled.
* **F1 Score**: The harmonic mean of precision and recall is 50.50%.
* **Success Class:** The success class is class 1 (canceled).
* **Success Probability:** The probability of success in predicting the positive class is 50%.

The model shows reasonable accuracy (66.67%), Class 1 (canceled) has a higher error rate (53.21%), indicating that the model struggles more with predicting cancellations. Specificity (78.01%) suggests the model performs relatively well in identifying non-canceled bookings. Sensitivity (46.79%) indicates room for improvement in identifying canceled bookings. Precision (54.84%) highlights the proportion of correctly predicted cancellations among all predicted cancellations. F1 score (50.50%) provides a balanced measure of precision and recall.

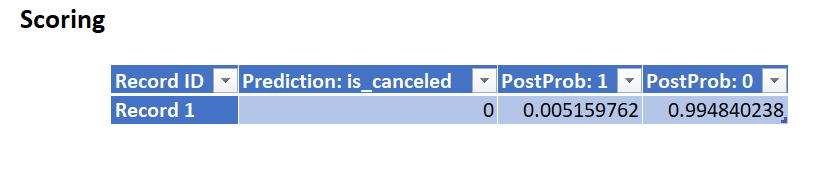
In summary, the model's performance is moderately successful, with areas for improvement, especially in correctly predicting cancellations (Class 1).

**classification of booking cancellation for a new customer with lead\_time = 216 and no previous cancellations.**

* In order to predict the classification for the above customer with lead\_time = 216, we have changed three variable values as mentioned below.

1. Lead\_time = 216
2. Previous\_cancellations =0
3. Is\_repeated\_guest = 0

* We have performed scoring based on the above data and predicted the booking cancellation as below.



**Result**: For a new customer with lead\_time = 216, hotel booking is not cancelled. The probability of booking not cancelled is 0.99 and the probability of booking cancelled is 0.005 which means 99 percent that booking will not be cancelled.

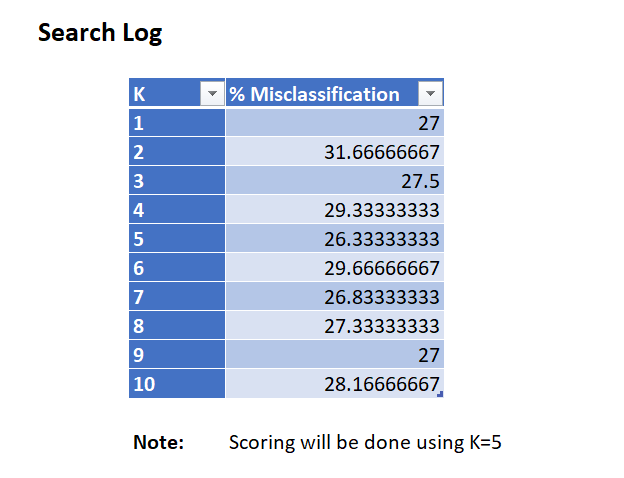
**TECHNIQUE 2 – KNN:**

K-nearest-neighbors algorithm is used for classification (of a categorical outcome) and prediction (of a numerical outcome). KNN is simple and no assumptions required about normal distribution.

* **Data Splitting:** The dataset is divided into training and validation set used for model training and a validation set used for model evaluation. The split ratio is 60 – 40 means we have divided 60 percentage of data into training set and 40 percentage of data into testing set.
* In this case, we have partitioned the data into 60 percent into training data and 40 percent data into validation dataset.
* **Feature Scaling:** Since KNN is distance-based, it is essential to scale the features to bring them to a similar scale. Standardization or normalization is commonly applied. So, in this case we are rescaling the data with standardization.
* **Choosing K:**

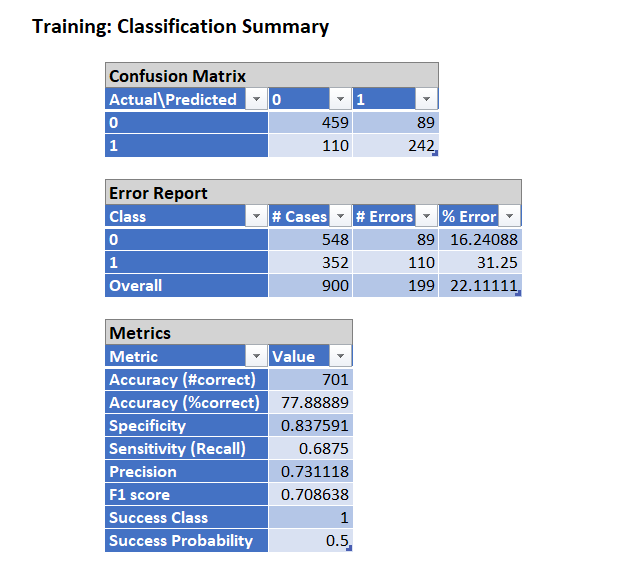
**Step1:** We have to identify the best k value for that we need to search for the best k value by keeping the K value maximum as 10.

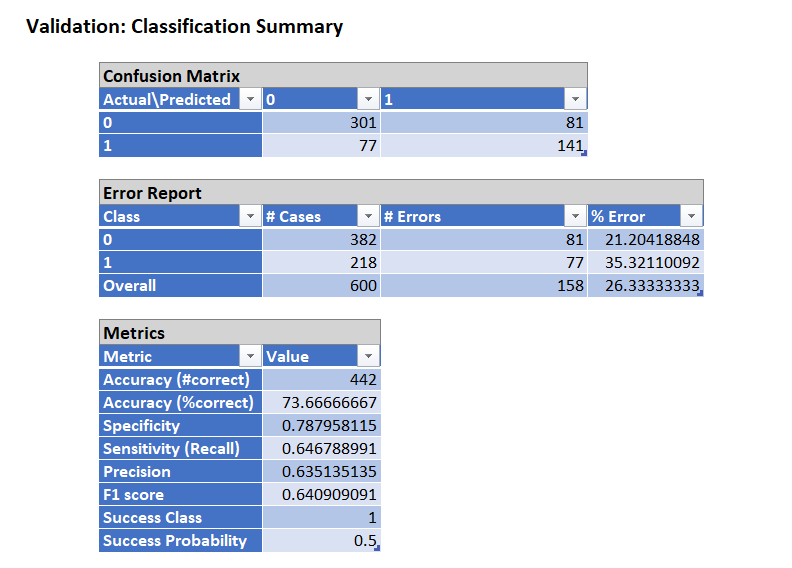
**Step 2:** Perform KNN algorithm in order to get the best K value. We will get the best K value as per the attached screenshot below.



**Step 3**: In this case, we got the best K value as 5. Now, we need to perform KNN with K value = 5. Output of KNN algorithm is given below.

**RESULT AND ANALYSIS:**





**INTERPRETATION FOR VALIDATION DATASET:**

**Confusion Matrix:**

**True Positives (TP):** 141 bookings were correctly predicted as canceled.

**True Negatives (TN):** 301 bookings were correctly predicted as not canceled.

**False Positives (FP):** 81 bookings were incorrectly predicted as canceled.

**False Negatives (FN):** 77 bookings were incorrectly predicted as not canceled.

* **Accuracy:** The model correctly predicted 73.67% of all cases.
* **Specificity:** The ability of the model to correctly identify non-canceled bookings is 78.80%.
* **Sensitivity (Recall):** The ability of the model to correctly identify canceled bookings is 64.68%.
* **Precision:** Among the predicted canceled bookings, 63.51% were actually canceled.
* **F1 Score:** The harmonic mean of precision and recall is 64.09%.
* **Success Class:** The success class is class 1 (canceled).
* **Success Probability:** The probability of success in predicting the positive class is 50%.

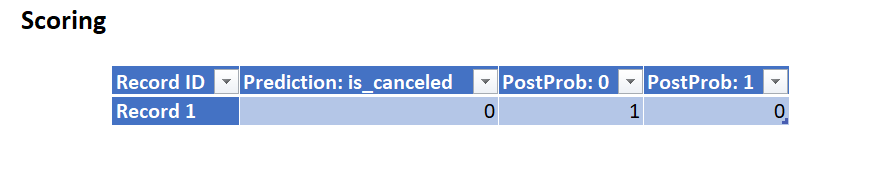
The KNN model demonstrates improved performance compared to the previous model (Logistic Regression). The accuracy has increased to 73.67%, indicating a higher overall correct prediction rate. Class-specific metrics show better performance in predicting cancellations (Class 1) with a sensitivity of 64.68%, compared to the previous model. Specificity remains high, indicating proficiency in identifying non-canceled bookings (Class 0). Precision and F1 score provide a balanced view of the model's ability to make accurate positive predictions.

**classification of booking cancellation for a new customer with lead\_time = 216 and no previous cancellations.**

* In order to predict the classification for the above customer with lead\_time = 216, we have changed three variable values as mentioned below.

1. Lead\_time = 216
2. Previous\_cancellations =0
3. Is\_repeated\_guest = 0

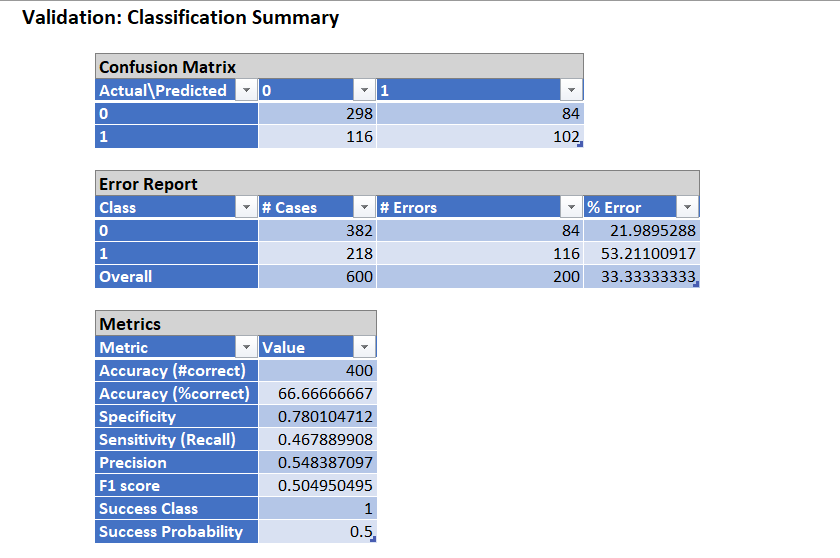
* We have performed scoring based on the above data and predicted the booking cancellation as below.



**Result**: For a new customer with lead\_time = 216, hotel booking is not cancelled. The probability of booking not cancelled is 1 and the probability of booking cancelled is 0 which means 100 percent that booking will not be cancelled.

**RESULTS:**

**LOGISTIC REGRESSION:**

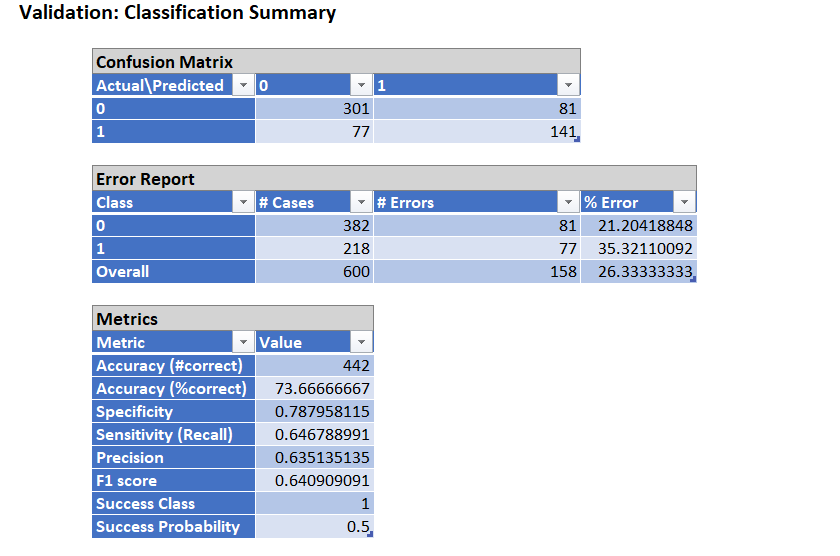
****

* **Accuracy:** The logistic regression model achieved an accuracy of 66.67%. This means it correctly predicted whether a booking would be canceled or not for approximately two-thirds of the cases.
* **Sensitivity (Recall):** Sensitivity or the ability to correctly identify cancellations

(Class 1) is 46.79%. This suggests the model has improvement in capturing instances where bookings are canceled.

* **Precision:** Precision represents the accuracy of positive predictions, is 54.84%. This indicates that among the predicted cancellations, 54.84% were indeed cancellations.
* **F1 Score:** The F1 score is a harmonic mean of precision and recall which is 50.50%, reflecting a balanced measure of the model's performance.
* The logistic regression model demonstrates moderate accuracy but struggles with correctly identifying cancellations (Class 1) as indicated by the lower sensitivity.

**KNN ALGORITHM:**

****

* **Accuracy:** The KNN model achieved a higher accuracy of 73.67%, suggesting improved overall performance compared to logistic regression.
* **Sensitivity (Recall):** Sensitivity for predicting cancellations (Class 1) has improved to 64.68%, indicating a better ability to capture instances where bookings are canceled.
* **Precision:** Precision remains strong at 63.51%, signifying accuracy in positive predictions.
* **F1 Score:** The F1 score for KNN is 64.09%, reflecting a balanced performance in precision and recall.
* The KNN model outperforms logistic regression with higher accuracy and improved sensitivity in identifying cancellations.

**COMPARISON:**

* **Accuracy:** KNN performs better in terms of overall accuracy, achieving a 7% higher accuracy compared to logistic regression.
* **Sensitivity:** KNN has higher sensitivity which indicates a better ability to identify instances of cancellations. This aligns with the business question's focus on predicting cancellations.
* **Precision:** Precision is comparable between the two models, KNN maintaining a slightly higher precision.
* **F1 Score:** KNN has a higher F1 score, indicating a more balanced performance in terms of precision and recall.

**SUMMARY:**

* Both models provide insights into the likelihood of hotel booking cancellations.
* Logistic Regression is achieving moderate accuracy may benefit from further refinement to improve sensitivity.
* On the other hand, KNN exhibits improved overall performance particularly in identifying instances of cancellations.

**RECOMMENDATIONS:**

* Consider further feature engineering or selection to enhance the predictive power of both models.
* Explore hyperparameter tuning for KNN to potentially improve its performance further.
* Evaluate additional models or ensemble techniques for improved accuracy.

**Excel Sheets attached below for Logistic Regression and KNN Analysis:**


**DATASET:**

****

**Recording Link:**

[**https://1drv.ms/f/s!AmUZlHy7usy9ghm\_DSctZxb087\_e?e=t0u3hs**](https://1drv.ms/f/s!AmUZlHy7usy9ghm_DSctZxb087_e?e=t0u3hs)