MBAN 6110 S: Data Science I
Group Project

A report submitted in partial fulfillment of the requirement for the degree of
Masters of Business Analytics (MBAN)

Schulich
School of Business
York University

YORK
UNIVERSITÉ
UNIVERSITY

Under the guidance of
Prof. Delina Ivanova

Submitted by –
Abdul Ghafar Qasemi – 220067401
Aimal Dastagirzada – 220088928
Mahin Bindra – 220089330
Qian Wang - 219346956
Yue Peng – 215140080
Yini Shi - 220117743

August 2023

I. EXECUTIVE SUMMARY

We have created a predictive model to solve the high last-minute hotel cancellation due to changes in place and scheduling conflicts. This model allows the hotel management to anticipate whether a potential customer will honor or cancel their reservation, enabling the hotel to implement effective strategies. These strategies aim to minimize the financial impact caused by last-minute cancellations, ensuring the hotel does not bear most of the losses.

To achieve this, we started by developing and evaluating several models to find out that the KNeighborsClassifier model excelled in predicting 'booking status' across the majority of the metrics – precision, F1-score and accuracy – making it the ideal choice for hotel management to anticipate and handle cancellations effectively.

Our model (**KNN**) demonstrated a **precision score of 89%**. Our analysis and modelling revealed that certain factors are essential in determining whether a customer would keep their reservation. Factors include lead time, the average price per room, the number of previous cancellations, and prior bookings not cancelled.

This report includes strategic recommendations we suggest for hotel management, which include –

1. Implementing higher cancellation fees for early bookings
2. Increasing cancellation charges as the arrival date approaches
3. Limiting free cancellation days to benefit the hotel financially
4. Offering incentives to encourage guests not to cancel, such as discounts and complimentary amenities
5. Identify successful room attributes and apply similar strategies to other rooms.

In conclusion, our project would suggest using the KNeighborsClassifier model to manage cancellations effectively. By implementing higher cancellation fees for early bookings and offering incentives, hotels can improve revenue management and enhance guest satisfaction, resulting in minimized financial losses.

II. INTRODUCTION

Online hotel reservation channels have revolutionized booking possibilities and customer behaviour. However, hotels are now grappling with financial losses due to customers cancelling their reservations or not showing up. This dataset reflects a similar issue where the hotel incurs financial losses due to customers not honouring their reservations.

In this report, we focus on analyzing the dataset to identify why customers fail to honour their reservations. To achieve this, we will develop a classification predictive model that considers various factors to forecast whether customers will keep their reservations.

By the end of the report, we will provide recommendations to the hotel management to mitigate losses, even in cases where customers book ahead of time and cancel on short notice before their arrival date. The aim is to implement strategies to ensure the hotel remains profitable despite potential reservation uncertainties.

I. Dataset Selection and Justification

For this research, we used a binary class open-sourced dataset called 'Hotel Reservations Dataset'. The dataset was obtained from Kaggle, and it contained approximately 36000 entries containing information on 18 different attributes such as lead time, average room per price, market segment type, number of previous cancellations, type of meal plan reserved etc. Although the dataset did not contain any missing values, we were to handle some outliers and transform the data using feature engineering before training the model. The dataset was first uploaded on 3rd Jan 2023 and last updated in February 2023. The information present in the dataset is from July 2017 to December 2018.

Based on an analysis conducted through a pair plot, the distinctions between cancelled and non-cancelled bookings seemed pronounced, with minimal overlaps between the two categories across most features. Thus, the utilization of classification models to analyze hotel booking history becomes an efficient approach, enabling the precise prediction of booking status.

II. Key Research Questions

- What are the main challenges encountered by hotels as a result of hotel reservation cancellations?
- What is the percentage of hotel cancellations and non-cancellations based on various market segments?
- How, if at all, do cancellation rates vary based on various factors (e.g., lead time, average room price etc.)?
- What strategies and recommendations can we devise to assist hotels in mitigating the challenges resulting from these cancellations?

III. Problem Statement

Many customers cancel their hotel reservations last minute due to changes in plans, scheduling conflicts, and the availability of better and cheaper rooms elsewhere. Since most bookings are made online, cancelling is mostly free of charge or only requires a small cancellation fee. However, this is an extreme cause of concern for the hotels as it results in diminishing revenue.

IV. Purpose

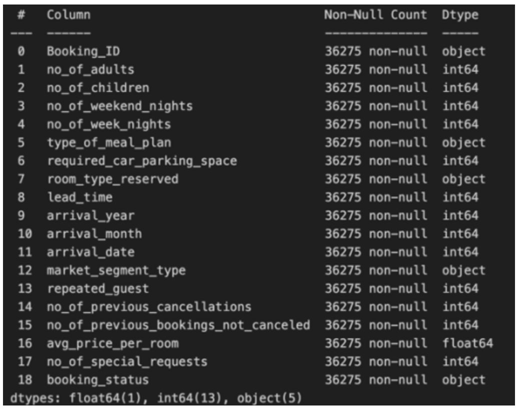
This research aims to assess whether a potential customer will cancel or honor their hotel reservation based on various metrics and devise strategies to assist the hotel management in avoiding diminishing revenues due to the low cost or free-of-charge cancellations.

III. ANALYSIS

I. Exploratory Data Analysis

Thorough data cleaning and handling of missing values

- *Missing values:* The dataset contains no missing values, and the data types are correct according to each feature's definition.



```
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Booking_ID                            36275 non-null  object
 1   no_of_adults                          36275 non-null  int64
 2   no_of_children                        36275 non-null  int64
 3   no_of_weekend_nights                  36275 non-null  int64
 4   no_of_week_nights                     36275 non-null  int64
 5   type_of_meal_plan                     36275 non-null  object
 6   required_car_parking_space            36275 non-null  int64
 7   room_type_reserved                    36275 non-null  object
 8   lead_time                             36275 non-null  int64
 9   arrival_year                          36275 non-null  int64
 10  arrival_month                         36275 non-null  int64
 11  arrival_date                          36275 non-null  int64
 12  market_segment_type                   36275 non-null  object
 13  repeated_guest                        36275 non-null  int64
 14  no_of_previous_cancellations          36275 non-null  int64
 15  no_of_previous_bookings_not_canceled  36275 non-null  int64
 16  avg_price_per_room                    36275 non-null  float64
 17  no_of_special_requests                36275 non-null  int64
 18  booking_status                        36275 non-null  object
dtypes: float64(1), int64(13), object(5)
```

Figure 3.1: Information about the dataset

- *Outliers:* 1728 of outliers are removed from the dataset, representing 4.7% of the data.
  - It appears that majority of the price rooms are within the 25% - 75% quartile. The outliers of "ave_price_per_room" that are not within this range were removed through Interquartile Range Method.
  - Most of the bookings have no child or only 1 child. Thus, any bookings contain more than 7 children are removed.
  - It appears some bookings contain odd high number of previous cancellations as the average number of previous cancellations is as low as 0.02. Thus, any bookings with more than 9 previous cancellations are removed.

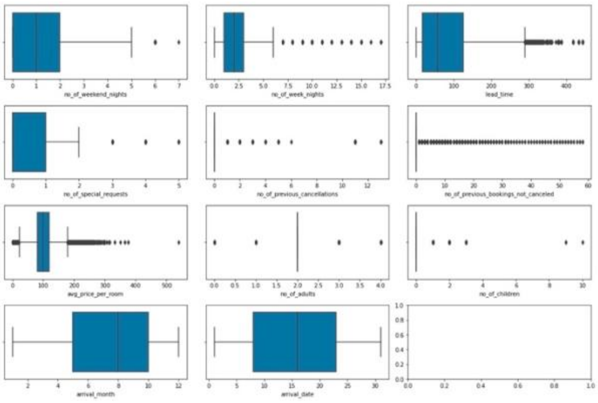


Figure 3.2: Boxplot showcasing the presence of outliers within different columns

Descriptive Statistics and Visualizations:

- *Average Price Per Room:* It appears that bookings with lower average price per room (less than 100) are less likely to be cancelled than those with higher average price per room. Additionally, bookings with average price per room in the range of 100 to 150 have the highest number of cancellations.

- *Lead Time:* It appears that bookings made further in advance tend to have a higher rate of cancellation. Specifically, bookings made within 0 - 50 days have a lower cancellation rate, while bookings made more than 50 days ahead have a higher cancellation rate. Therefore, it may be useful for the hotel to offer discounts or other incentives to encourage guests to book closer to their travel dates, as this could help reduce the cancellation rate and increase overall revenue.

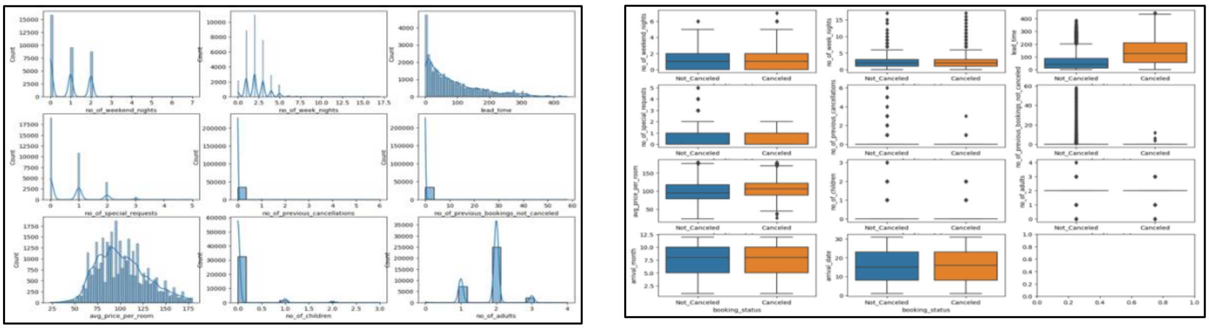
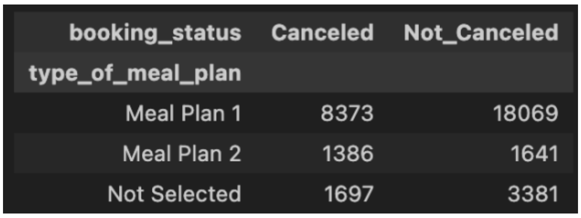


Figure 3.3: Histplot and Boxplot for all the continuous variables

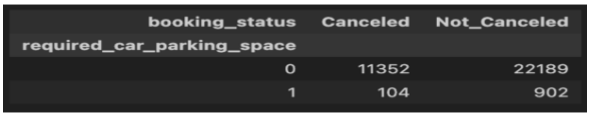
- *Meal Plan:* It appears that majority of the customers (over 73%) preferred “Meal Plan 1” than the other selections, with around 31% of customers who choose Meal Plan 1 canceled the booking. The second preferred choice was “Not Selected” (8%) with around 34% of customers who choose “Not Selected” canceled the booking. The least popular meal plan is “Meal Plan 2” with highly negligible percentage, and the cancellation rate of the customers who choose Meal Plan 2 is nearly 50%. Thus, the hotel should focus on meal plan 1 as it drawn more customers and produced the lowest cancellation rate. And disregard the Meal Plan 2.



booking_status	Canceled	Not_Canceled
type_of_meal_plan		
Meal Plan 1	8373	18069
Meal Plan 2	1386	1641
Not Selected	1697	3381

Figure 3.4: Segregation of type of meal plan by booking status

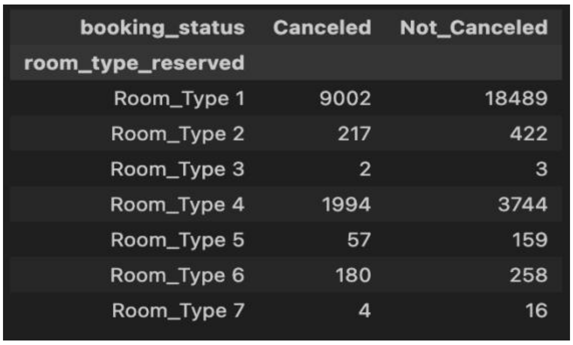
- *Car Parking:* Most of the booking did not require car parking. Hotel received about 33% of cancellation from bookings without a car parking and 10% of cancellation from those with 1 required parking.



booking_status	Canceled	Not_Canceled
required_car_parking_space		
0	11352	22189
1	104	902

Figure 3.5: Segregation of parking space requirement by booking status

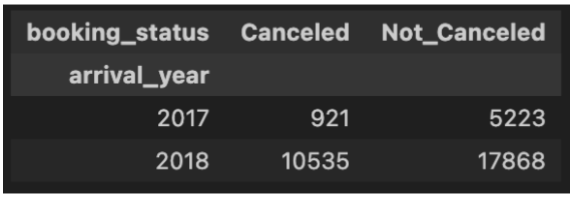
- *Room Type:* "Room Type 1" witnessed the largest intake of customers (75% of the bookings), with 32% of those booked Room Type 1 ended up canceled. Whereas, except for Room Type 4, the rest of the room types takes negligible percentage.



booking_status	Canceled	Not_Canceled
room_type_reserved		
Room_Type 1	9002	18489
Room_Type 2	217	422
Room_Type 3	2	3
Room_Type 4	1994	3744
Room_Type 5	57	159
Room_Type 6	180	258
Room_Type 7	4	16

Figure 3.6: Segregation of room type reserved by booking status

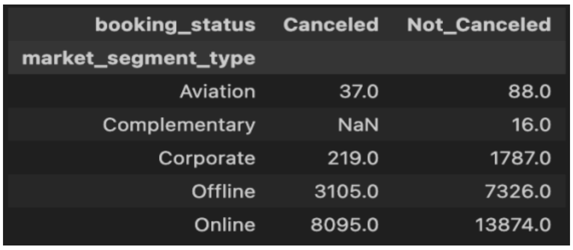
- *Arrival Year:* In both 2017 and 2018, the number of bookings that were not cancelled are more than those were canceled. And the total bookings in 2018 is more than that in 2017, which means hotel managed to attract more customers within a year, indicating a sign of positive progress. However, the cancellation rate in 2018 was 37% which is 2.5 times of the cancellation rate in 2017 (14.9%), indicating that increased number of customers resulted in higher chances of cancellations.



booking_status	Canceled	Not_Canceled
arrival_year		
2017	921	5223
2018	10535	17868

Figure 3.7: Segregation of arrival year by booking status

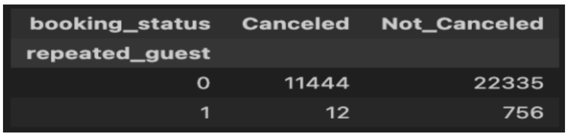
- *Market Segment Type:* The majority of the bookings were made online, with a total number of 22669, and the highest cancellation rate (36.8%) occurred for online bookings. Offline booking was the second most popular selection, with 10431 bookings made offline, and the cancellation rate is at 29.7%. Interestingly, Complementary order had very low number of booking but there was no any cancellation through this booking. And Corporate booking also had low cancellations rate, even though the number of bookings was low.



booking_status	Canceled	Not_Canceled
market_segment_type		
Aviation	37.0	88.0
Complementary	NaN	16.0
Corporate	219.0	1787.0
Offline	3105.0	7326.0
Online	8095.0	13874.0

Figure 3.8: Segregation of market segment type by booking status

- *Repeated Guest:* Nearly all the bookings (93%) were by new customer, showing higher cancellation rate (33.8%). However, the cancellation rate from repeated guest shows much lower cancellation rate, which is only 1.6%.



booking_status	Canceled	Not_Canceled
repeated_guest		
0	11444	22335
1	12	756

Figure 3.9: Segregation of repeated guest status by booking status

II. Feature Engineering

In our Exploratory Data Analysis, we discovered several variables that exhibit strong predictive potential for our target/dependent variable. To initiate the feature engineering process, we took the following steps.

1) Making Dummy variable:

We started by making dummy variables for all the categorical variables that showed strong predictive potential. We created these dummy variables to convert categorical data into a numerical format that can be used as input for machine learning algorithms. The process of creating these dummy variables was accomplished using the 'get_dummies' function from the pandas DataFrame. Creating these dummies ensures that the models can use categorical information effectively to make accurate predictions while keeping the results easy to understand.

2) New Data Frame:

After creating dummy variables for categorical data, a new DataFrame named 'new_df' was formed by concatenating the continuous variables identified as strong predictors during exploratory data analysis, along with the previously generated dummy variables. This new DataFrame 'new_df' now holds both continuous and categorical predictors, ready for further analysis and modeling.

3) Custom Function:

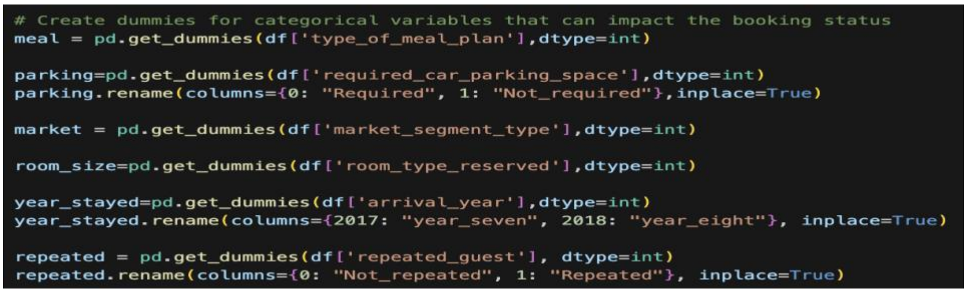
After completing feature engineering, the next step involved creating a custom function to convert our target variable into binary format. This transformation allowed us to represent the target variable as either a 0 or 1, making it suitable for binary classification tasks.

4) Data Transformation Pipelines for Categorical and Numerical Features

We implemented two pipelines to preprocess our data for machine learning.

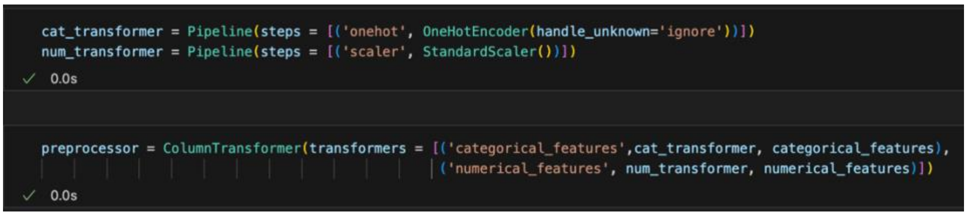
- Categorical Feature Pipeline (cat_transformer): Utilizing 'OneHotEncoder', we converted categorical variables into a one-hot encoded format. This transformation enables seamless integration of categorical data into ML models.
- Numerical Feature Pipeline (num_transformer): By employing 'StandardScaler', we standardized numerical features, preventing any undue dominance during model training. These pipelines ensure our data is well-prepared for accurate model training and analysis, forming a strong foundation for our machine learning project.

We use these pipelines to prepare our data for machine learning. The 'cat_transformer' converts categorical variables to one-hot encoded format, and the 'num_transformer' standardizes numerical features. This ensures data is ready for accurate model training and analysis in our machine learning project.



```
# Create dummies for categorical variables that can impact the booking status
meal = pd.get_dummies(df['type_of_meal_plan'],dtype=int)

parking=pd.get_dummies(df['required_car_parking_space'],dtype=int)
parking.rename(columns={0: "Required", 1: "Not_required"},inplace=True)

market = pd.get_dummies(df['market_segment_type'],dtype=int)

room_size=pd.get_dummies(df['room_type_reserved'],dtype=int)

year_stayed=pd.get_dummies(df['arrival_year'],dtype=int)
year_stayed.rename(columns={2017: "year_seven", 2018: "year_eight"}, inplace=True)

repeated = pd.get_dummies(df['repeated_guest'], dtype=int)
repeated.rename(columns={0: "Not_repeated", 1: "Repeated"}, inplace=True)
```

Figure 3.10: Dummy Variables



```
cat_transformer = Pipeline(steps = [('onehot', OneHotEncoder(handle_unknown='ignore'))])
num_transformer = Pipeline(steps = [('scaler', StandardScaler())])
```

✓ 0.0s

```
preprocessor = ColumnTransformer(transformers = [('categorical_features',cat_transformer, categorical_features),
                                                 ('numerical_features', num_transformer, numerical_features)])
```

✓ 0.0s

Figure 3.11: Pipelines for Categorical and Numerical Features

III. Model Development

Post-assessing the relationship between multiple inputs and the target variables, and feature engineering the dataset, we were ready to build our model to predict whether the potential customers with honor or cancel their bookings. The modelling process went through the following steps –

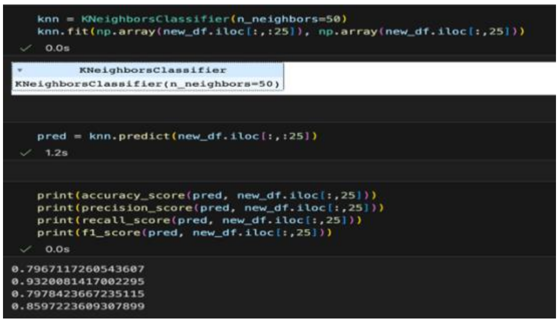
1) Identifying dependent variables

Recognizing which variables are the most important predictors of the ‘booking_status’, the target variable and exclude all independent variables from the updated data frame (new_df). The variables identified as most important predictors in are our case are –

- Lead time – number of days before the arrival date the booking was made
- No. of previous cancellations – number of previous cancelled by customer
- No. of previous bookings not cancelled – number of bookings not cancelled
- Average price per room – average price per day of the room
- Type of meal plan – type of meal plan included in the booking
- Market segment type – how the booking was made
- Room type reserved – the type of room reserved at the time of booking
- Arrival year – year of arrival
- Repeated guest – whether the guest has previously stayed at the hotel
- Required parking space –whether a car parking space is required

2) Preliminary testing on entire dataset

Before partitioning the dataset into train and test split and developing models, we conducted a preliminary test by creating a KNN model and predicting values based on complete data. Using this technique, we wanted to check the accuracy of a classification model based on the raw dataset to assess whether we can proceed with developing predictive models or if the dataset still requires transformation.



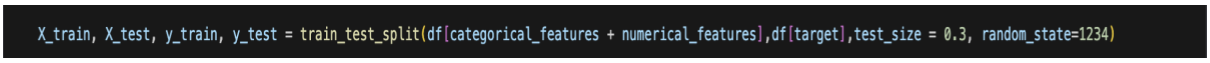
```
knn = KNeighborsClassifier(n_neighbors=50)
knn.fit(np.array(new_df.iloc[:,:25]), np.array(new_df.iloc[:,25]))
✓ 0.0s

▾ KNeighborsClassifier
KNeighborsClassifier(n_neighbors=50)

pred = knn.predict(new_df.iloc[:,:25])
✓ 1.2s

print(accuracy_score(pred, new_df.iloc[:,25]))
print(precision_score(pred, new_df.iloc[:,25]))
print(recall_score(pred, new_df.iloc[:,25]))
print(f1_score(pred, new_df.iloc[:,25]))
✓ 0.0s

0.7967117260543607
0.9320081417002295
0.7978423667235115
0.8597223609307899
```

Figure 3.12: Preliminary model creation – KNeighborsClassifier

Based on the results above, we can proceed with developing and evaluating different classification models to predict whether the customers will honor or cancel their reservations.

3) Partition the data into train and test splits

After preparing the dataset, we partitioned it into a train and test split of 70:30. We used the train set to train the model based on current data, and the test dataset was reserved to evaluate the model's performance.

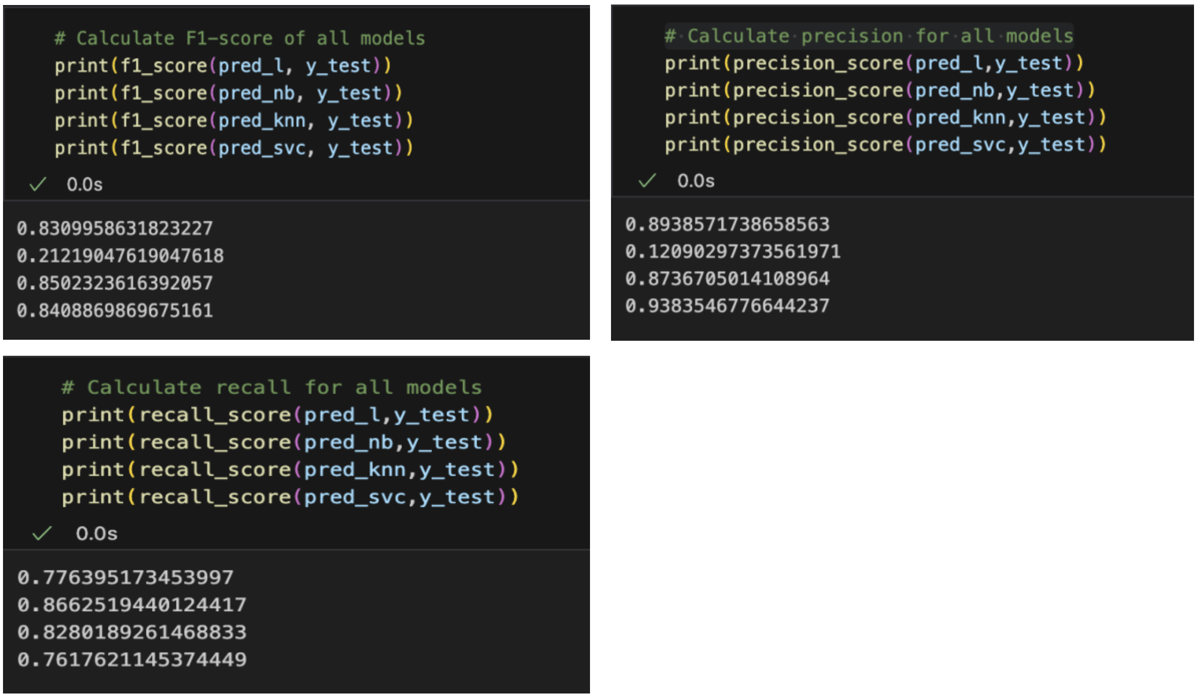


```
X_train, X_test, y_train, y_test = train_test_split(df[categorical_features + numerical_features],df[target],test_size = 0.3, random_state=1234)
```

Figure 3.13: Partitioning data into train and test split

4) Develop and evaluate different models

We developed multiple models and fitted them to ultimately choose the best model based on its prediction efficiency. The ML algorithms implemented in this case were Logistic Regression, K-Nearest Neighbor, Gaussian Naïve Bayes and Support Vector Machines. We compared the results of the models mentioned above on the following metrics – F-measure, Precision and Recall.



```
# Calculate F1-score of all models
print(f1_score(pred_l, y_test))
print(f1_score(pred_nb, y_test))
print(f1_score(pred_knn, y_test))
print(f1_score(pred_svc, y_test))
✓ 0.0s

0.8309958631823227
0.21219047619047618
0.8502323616392057
0.8408869869675161
```

```
# Calculate precision for all models
print(precision_score(pred_l,y_test))
print(precision_score(pred_nb,y_test))
print(precision_score(pred_knn,y_test))
print(precision_score(pred_svc,y_test))
✓ 0.0s

0.8938571738658563
0.12090297373561971
0.8736705014108964
0.9383546776644237
```

```
# Calculate recall for all models
print(recall_score(pred_l,y_test))
print(recall_score(pred_nb,y_test))
print(recall_score(pred_knn,y_test))
print(recall_score(pred_svc,y_test))
✓ 0.0s

0.776395173453997
0.8662519440124417
0.8280189261468833
0.7617621145374449
```

Figure 3.14: Success metrics values of all models (Logistics regression, KNN, SVC and NB)

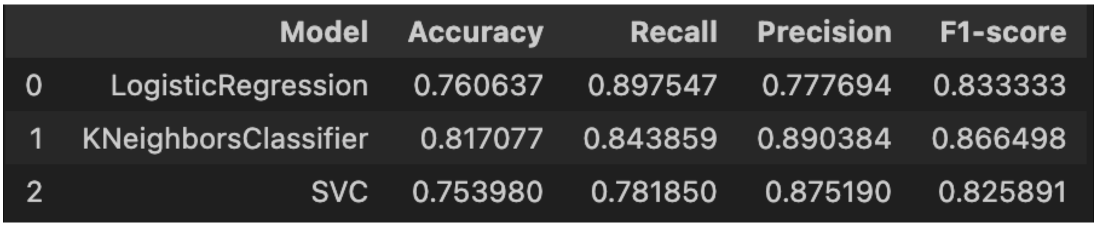
Based on the results above, we can deduce that SVC, Logistic Regression, and KNN models perform better on different metrics. However, precision is the most vital metric to measure the success of our research. Accurately predicting whether the customer will cancel is essential to avoid losses and operational challenges due to last-minute cancellations. Prioritizing precision is also crucial to minimize false positives and ensure correct predictions. Therefore, based on the precision score, SVC performs the best initially. However, we will fine-tune all three models and evaluate their results to make an informed decision.

5) Fine-tune the model to improve the efficiency

Fine-tuning the models and hyperparameters ensures optimal model performance and prevents overfitting the pre-trained model. We perform randomized search cross-validation to fine-tune and optimize the KNN, Logistic Regression and SVC models for precision. We adopted the 'RandomizedSearchCV' method over 'GridSearchCV'; even though it doesn't find the absolute best hyperparameters, it finds relatively good hyperparameters in a decent time frame. 'RandomizedSearchCV' is also more cost-effective, scalable, and less prone to overfitting than 'GridSearchCV'.
Upon fine-tuning, we observed an improved precision for the KNN model; however, the precision for SVC and Logistics regression models reduced slightly.

IV. Model Evaluation

After fine-tuning all the models using randomized search cross-validation, we evaluated all three models based on the following metrics – accuracy, precision, recall and f1-score, keeping in mind that the model having a higher precision would be given precedence considering the business context and implications.



	Model	Accuracy	Recall	Precision	F1-score
0	LogisticRegression	0.760637	0.897547	0.777694	0.833333
1	KNeighborsClassifier	0.817077	0.843859	0.890384	0.866498
2	SVC	0.753980	0.781850	0.875190	0.825891

Figure 3.15: Collective results to evaluate all models

Based on the results above for all three models, it is evident that the KNeighborsClassifier model performs the best in predicting the 'booking_status' on all metrics. Therefore, we selected the KNeighborsClassifier as our final model in predicting whether a customer will cancel or honor their reservation based on historical data. Its interpretability and ease of implementation made it an ideal choice for hotel management to use for prediction and decision-making.
In conclusion, the model development process involved the implementation of various algorithms, hyperparameter tuning, and cross-validation to ensure optimal model performance. The final selection of the KNeighborsClassifier model was justified by its superior performance, robustness, and interpretability, making it a valuable tool for hotel management to predict and mitigate free-of-charge cancellations.

IV. CONCLUSIONS

Since the advent of online hotel booking systems, customers have had to flexibility to book a hotel well in advance and cancel anytime due to last-minute changes, scheduling conflicts etc., at a low cancellation fee or free of charge. Although highly beneficial for the customers, it negatively impacts hotel revenues. Hence, it is crucial to develop predictive models that can prognosticate whether a potential customer will honor or cancel their hotel reservation. Through this research, we evaluated various classification models that can help achieve our goal. The outcomes showed that the KNeighborClassifier model outperformed all other models [Logistics Regression, Gaussian Naïve Bayes and Support Vector Machines] in predicting the 'booking status' of a potential customer based on historical data, on most success metrics, especially Precision as it's a vital success criterion. The KNeighborClassifier showcased a precision score of 89%, an accuracy of 81.7%, a recall score of 84.3% and an F1 score of 86.6%.

The results we achieved were satisfactory qualitatively. There is room for improvement in the technique used and further optimizing the model.

V. RECOMMENDATIONS

To avoid hotels from incurring financial losses due to last-minute cancellations and no-shows, we recommend the hotel management follow the below-mentioned implementation plan.

As the cancellation of hotel reservations depends on various factors such as lead time, average price per room, type of room etc., the hotel management should rigorously analyze cancellation patterns to identify trends. This approach can help them mitigate these cancellations and revenue losses that come with it by implementing targeted strategies.

1. Considering that lead time, i.e., the number of days before the arrival date the customer made the booking, is a vital predictor of whether the customer will honour their reservation or not, we suggest –
    - **Enforce a relatively higher cancellation fee** for customers booking way in advance, as the likelihood of people cancelling is higher when they have enough time to look for cheaper and better options.
        - Although doing so will not force the customers to honour their reservations, it will allow hotels to mitigate currently faced losses to a certain extent, preparing them well before a potential cancellation.
    - **Progressively increase cancellation fees as the lead time reduces** rather than having a consistent cost, making customers reconsider rescinding their reservation for fear of losing money.
        - The hotel management can roll out a plan where they increase the % of cancellation fees as the arrival date comes closer. For example –
            - Charge the entire room cost if a customer cancels two weeks before their arrival date.

        - Enforce a 75% cancellation fee if the customer cancels a month to 45 days before the arrival date.
        - Impose a 50% fee for cancelling 2-3 months before the booking date.
        - Lastly, levy a 30% fee of the total room cost if cancelled more than four months before the arrival date.
    - **Limiting the number of days when a customer can cancel free of charge** can allow customers to cancel a reservation in case of sudden changes in plans or mishaps without charge—also giving the hoteliers a financial advantage compared to the current market scenario if a customer exceeds the allotted limit.

2. Incentives for non-cancellation of bookings can encourage customers to honour their reservations, reducing the likelihood of cancellations.
    - Looking at the number of non-repeated guests who cancelled their bookings, although less than non-repeated guests who honoured their reservation, is still a high number. **Incentivizing these guests** by offering discount coupons, complimentary amenities, and special perks can make them commit to their reservations, reducing the revenue lost by hotels.
    - Since 'Room Type 1' experiences the most significant intake of guests, the hotel management can assess and identify attributes that contribute to such a large number of bookings, especially compared to 'Room Type 2' and 'Room Type 3', which have similar average prices. Using the results from the analysis, the hotel management can implement similar strategies to encourage more customers to book other rooms as well.

By adopting these recommendations, hotels should be able to avoid and reduce the number of last-minute cancellations due to changes in plans, scheduling conflicts etc., ultimately enhancing their revenue management strategies and improving guest satisfaction. However, the hotel management must conduct an extensive analysis before planning to avoid any reduction in overall footfall.

VI. APPENDIX

References

Raza, Ahsan. “Hotel Reservations Dataset.” *Kaggle*, 2023, https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset. Accessed 6 August 2023.

List of Figures