RESERVATION CANCELLATION PREDICTION

**1. Introduction**

**1.1 Overview**

As you can imagine, the cancellation rate for bookings in the online booking industry is quite high. Once the reservation has been cancelled, there is almost nothing to be done. This creates discomfort for many institutions and creates a desire to take precautions. Therefore, predicting reservations that can be cancelled and preventing these cancellations will create a surplus value for the institutions. In this article, I will try to explain how future cancelled reservations can be predicted in advance by machine learning methods. Customer behaviour and booking possibilities have been radically changed by online hotel reservation channels. Cancellations or no-shows cause a significant number of hotel reservations to be cancelled. Cancellations can be caused by a variety of factors, such as scheduling conflicts, changes in plans, etc. In many cases, this is made easier by the possibility of doing so free or at a low cost, which is beneficial for hotel guests but less desirable and possibly revenue-diminishing for hotels. It has 32 variables which include reservation and arrival date, length of stay, cancelled or not, the number of adults, children, or babies, the number of available parking spaces, how many special guests, companies, and agents pushed the reservation, etc. There are many aspects considered when choosing a hotel. The main idea is to find the appropriate prediction model for predicting hotel booking cancellations which finds the finest explaining variables for customer cancellations. This Hotel Booking Cancellation model can be useful not only for the vacationer but for the hotels’ owners.

**1.2 Purpose**

Hotel industry is one industry that can utilize data and machine learning in helping management to increase revenue. one of them by using machine learning as a tool to take preventive action against the possibility of a consumer to cancel their bookings**.**

So in this project I was using booking data from a hotel in Portugal that is used to take insight and design a simple machine learning model.

Before you initiate data understanding or analysis or feed your data to a machine learning algorithm, the first step is to clean your data and make it in a proper form (satisfying all the conditions). It is important to understand the type of data, patterns, and correlations present in your data. Here, is the Exploratory Data Analysis comes forward, The idea of getting to know your data in depth is called Exploratory Data Analysis.

After collecting data we have to import the necessary libraries to build machine learning models. Numpy, Pandas, Matplotlib, Seaborn are the known libraries used in the machine learning model.

**2. Literature Survey**

**2.1 Existing Problem**

The existing approaches or methods to solve this problem are:

* Exploratory Data Analysis
* Data Pre-Processing
* Model Building
* Model Comparison

The mindset behind this is to survey and understand the data before building a model. In today’s world, Data scientists and data analysts have to go through the most time in Data Wrangling and Exploratory Data Analysis (EDA) than building the models. In the EDA, using some statistical graphs and other visualization techniques, Data Scientists and Analysts try to find different patterns, missing data, relations, outliers, and anomalies in the data. EDA helps to ask the analytical questions and visualize the answer. It varies from person to person. Commonly the questions are related to independent variables and the target variable.

**2.2 Proposed Solution**

After collecting data we have to import the necessary libraries to build machine learning models. Numpy, Pandas, Matplotlib, Seaborn are the known libraries used in the machine learning model.

Numpy: NumPy is a Numerical Python Library that helps perform mathematical operations.

Pandas: Panda is an open-source library that helps understand relational or labelled data.

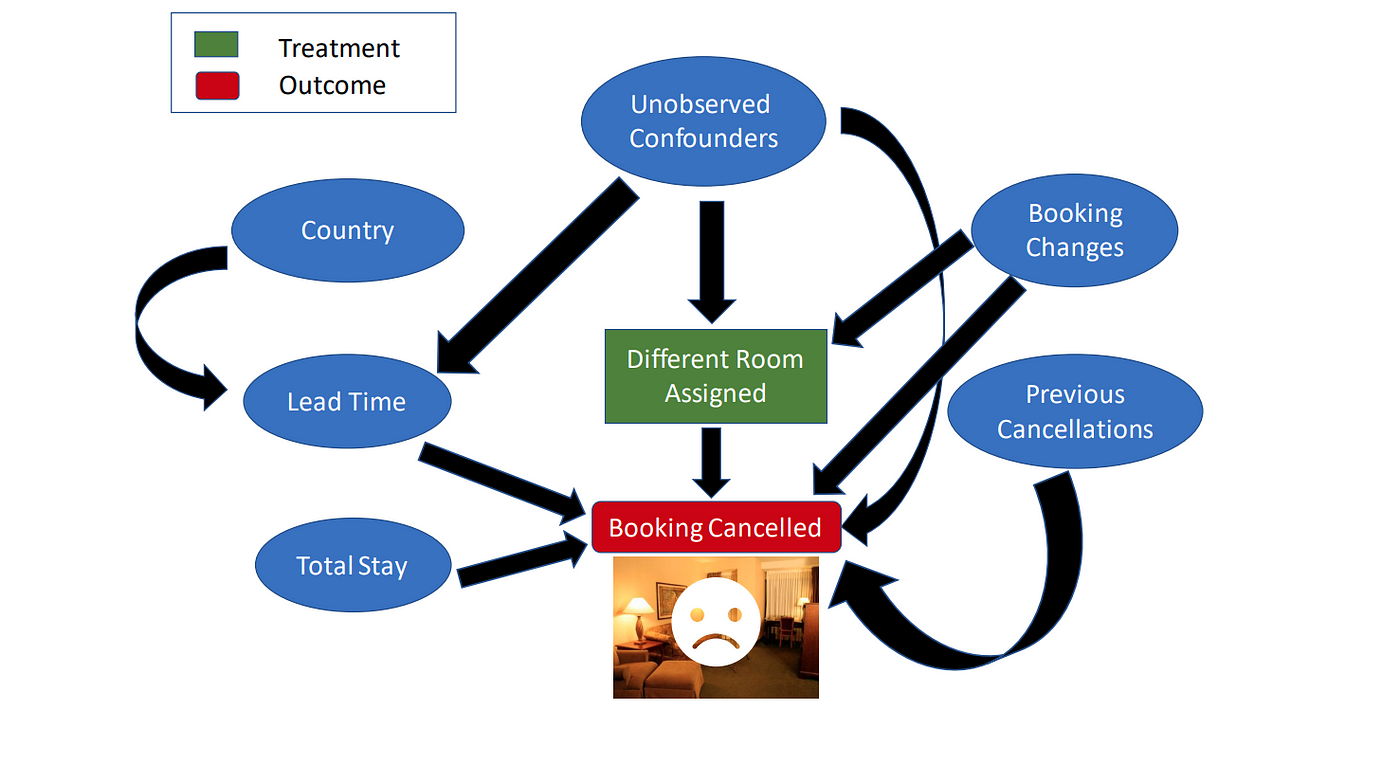
Matplotlib: Matplotlib is a Python visualization library that helps visualize 2D array plots.

Seaborn: Seaborn is a data visualization library built on top of matplotlib.

**3. Theoretical Analysis**

**3.1 Block Diagram**

Diagrammatic overview of the project



**3.2 Hardware/ Software Designing**

Hardware and software requirements of the project

Hardware: PC

Software: Anaconda (Jupyter)

Anaconda is package manager. Jupyter is a presentation layer. Anaconda tries to solve the dependency hell in python—where different projects have different dependency versions—so as to not make different project dependencies require different versions, which may interfere with each other.

The Jupyter Notebook application allows you to create and edit documents that display the input and output of a Python or R language script. Once saved, you can share these files with others. NOTE: Python and R language are included by default, but with customization, Notebook can run several other kernel environments.

**4. Experimental Investigations**

**Investigation made while working on the solution**

Data collection is the process of finding data from different public sites, or one can buy from private organizations and load data into our system. Kaggle and Github are familiar and known sites that provide free data. Data cleaning;

Real-world raw data is often incomplete, inconsistent, and lacking in certain behaviours or trends. The raw data contain many errors. So, once collected, the next step machine learning pipeline is to clean data which refers to removing unwanted data.

Some steps which are used to clean data are:

* Remove missing values, outliers, and unnecessary rows/ columns.
* Check and impute null values.
* Check Imbalanced data.
* Re-indexing and reformatting our data.

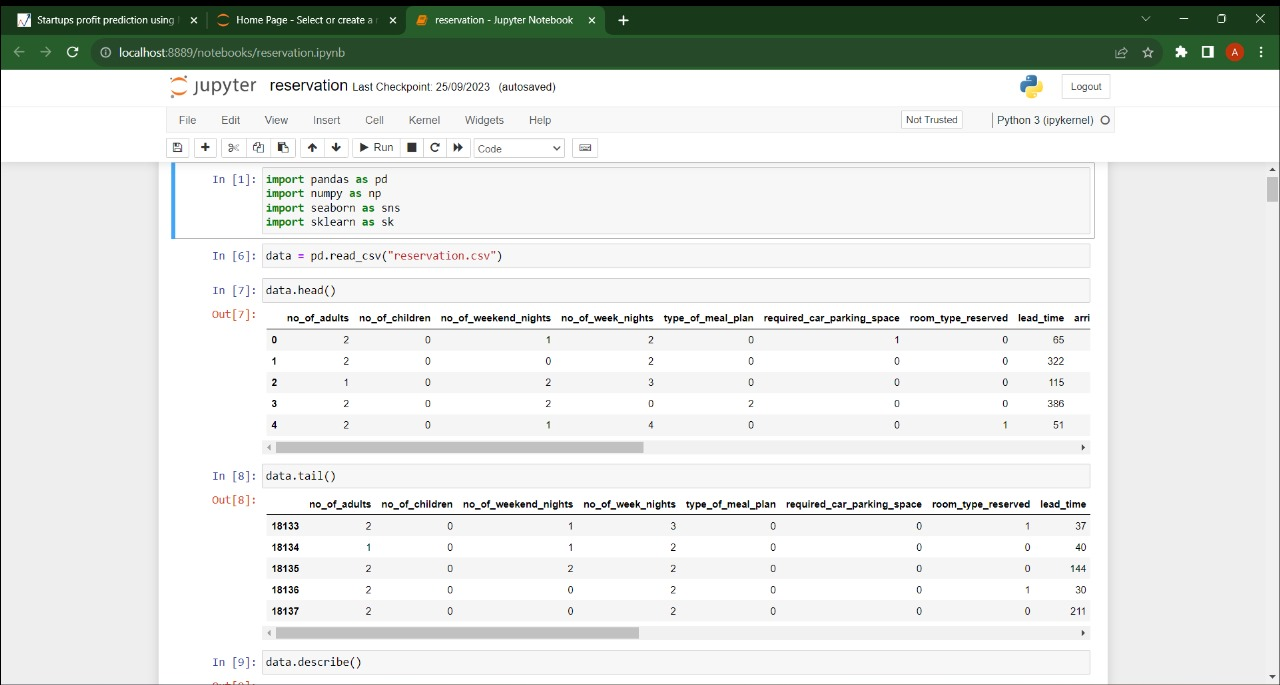
In the univariate analysis, we consider only a single variable that can be numerical or categorical. For numerical continue variables, we can use a histogram or scatter plot, and for categorical data, we commonly preferred bar plots or pie charts. In Numeric-Numeric analysis, we compare both numeric variables. The scatter plots, pair plots, and correlation matrix compare two numeric columns. In numeric-categorical analysis, one variable is a numeric type and the other is a categorical variable.

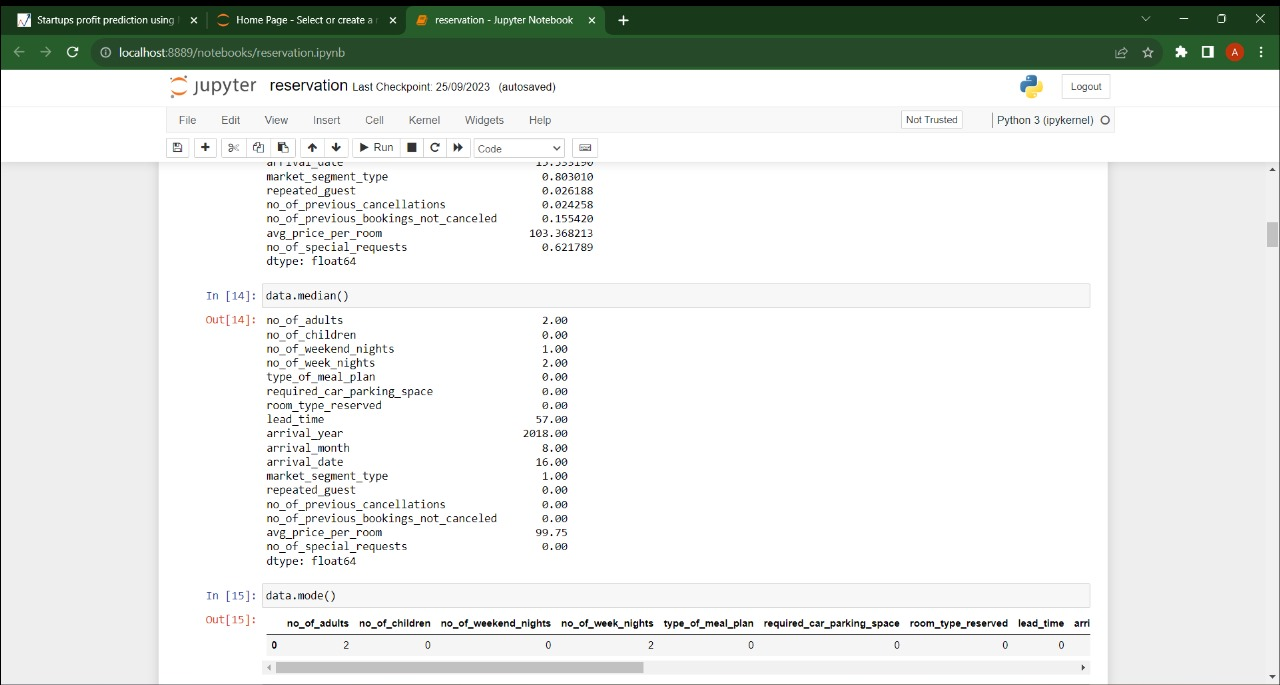
We can use group b The chi-square test determines the association between categorical variables. The Chi-square test calculates based on the difference between expected frequencies and the observed frequencies in one or more categories of the frequency table.

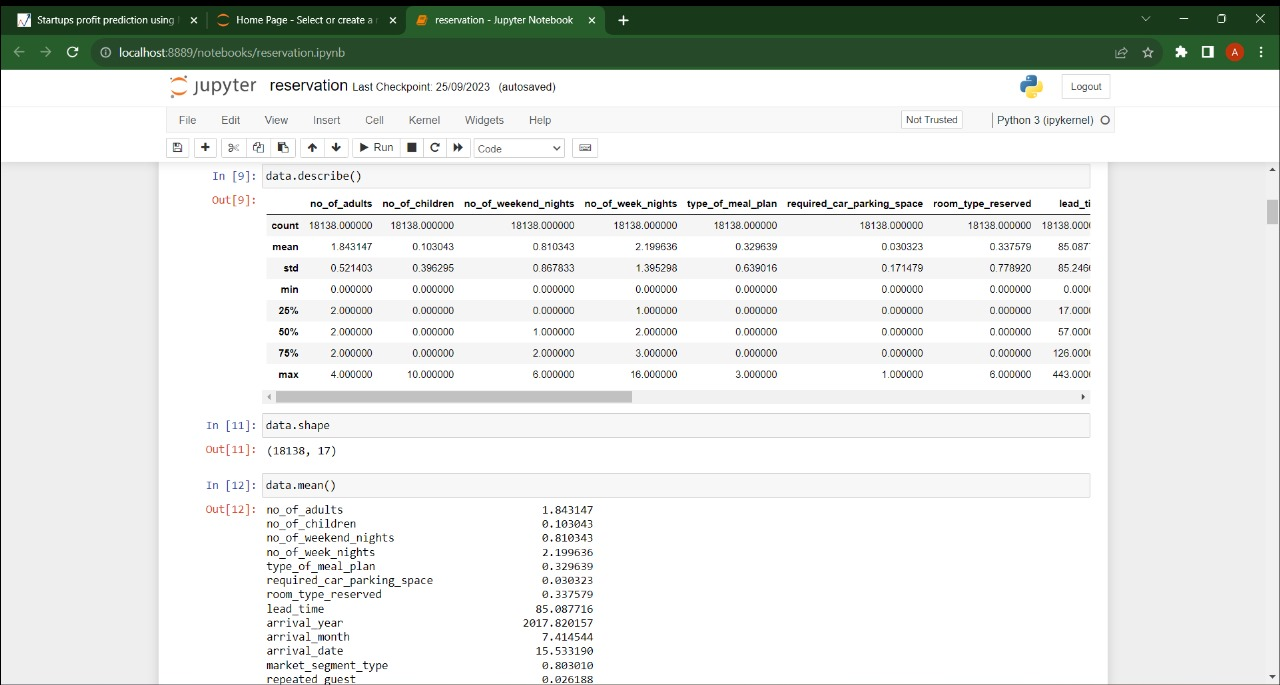
The Zero probability indicates a complete dependency between two categorical variables. The One probability indicates two categorical variables are completely independent. y function or box plot to perform numeric-categorical analysis.

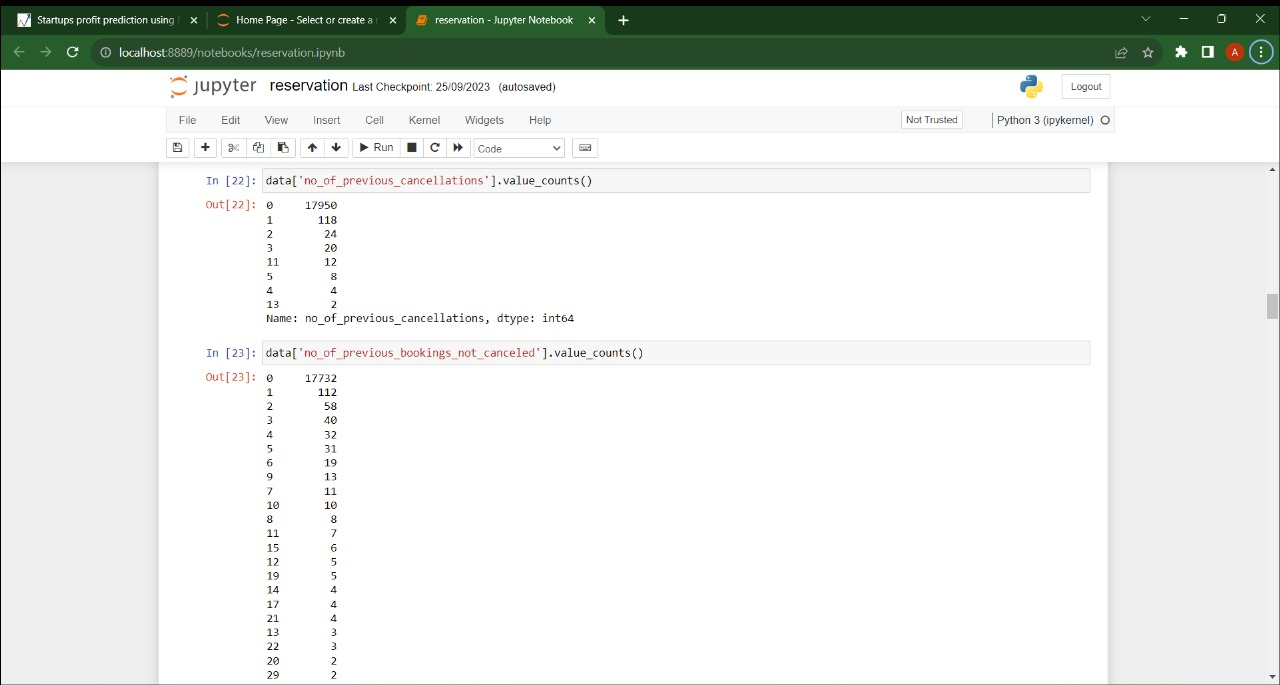
**5. RESULT**

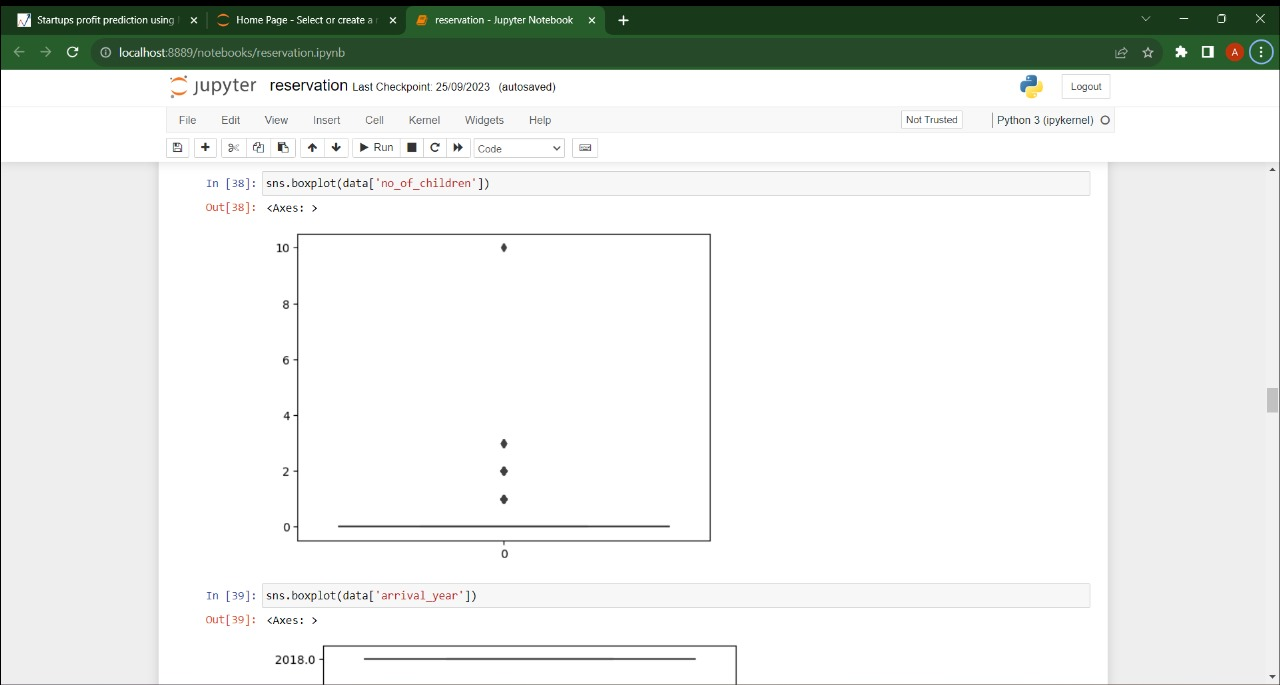
**Output of the project**

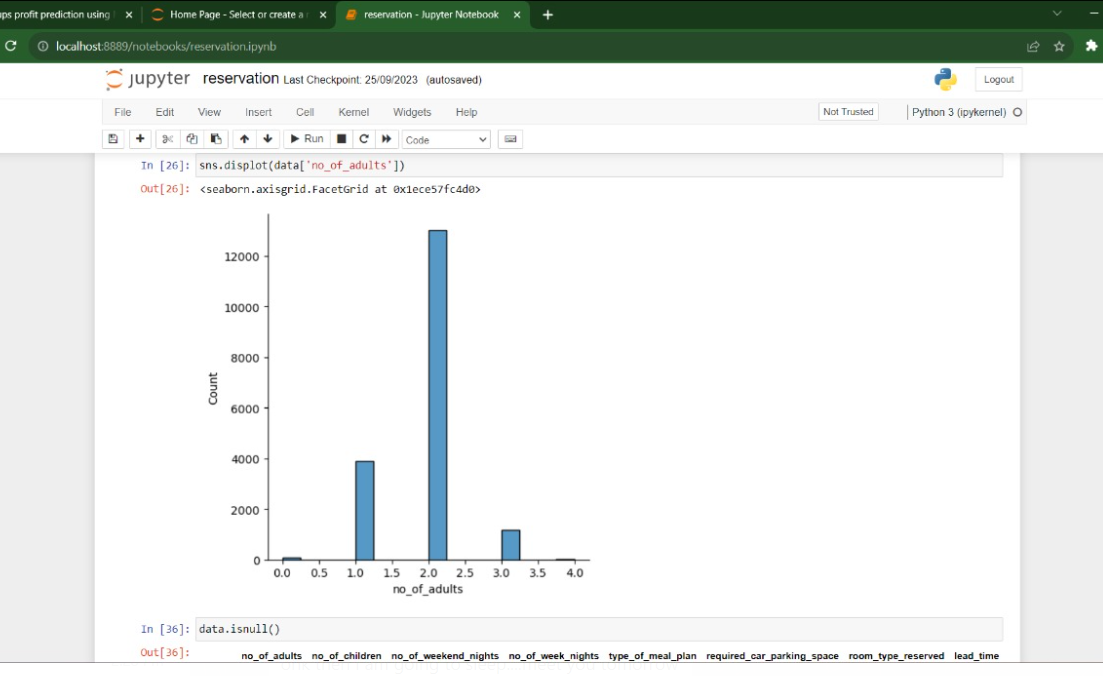
****

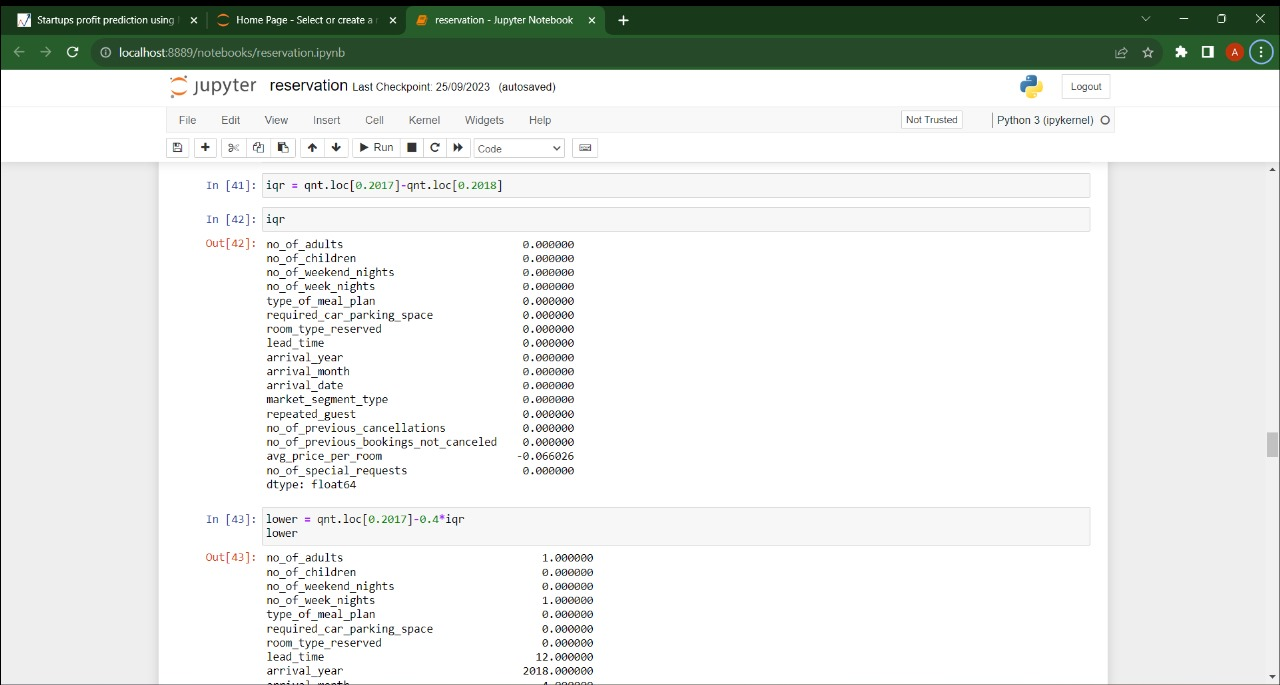
****

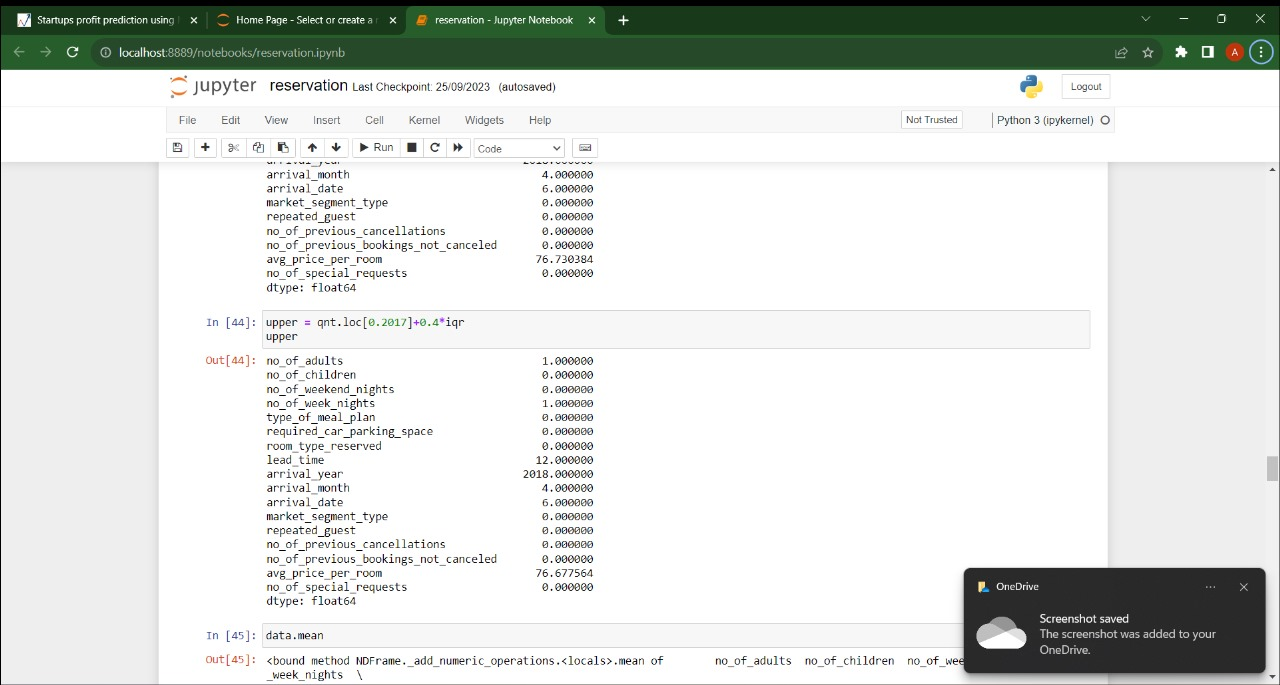
****

****









**6. ADVANTAGES AND DISADVANTAGES**

**Advantages:**

Optimized Resource Allocation: Hotels and other businesses can optimize their resources such as staff, inventory, and facilities based on accurate predictions, reducing wastage and increasing efficiency.

Improved Customer Service: Knowing about cancellations in advance allows businesses to offer vacant slots to other customers, enhancing overall customer satisfaction and experience.

Revenue Management: Predicting cancellations can aid in dynamic pricing strategies, maximizing revenue by adjusting prices based on demand fluctuations.

Data-Driven Decisions: Data from cancellation predictions can inform strategic decisions, helping businesses adapt their marketing and operational strategies accordingly.

Customer Retention: By offering personalized incentives or alternatives to potential cancellers, businesses can retain customers who might have cancelled otherwise.

**Disadvantages:**

Privacy Concerns: Predictive models often require access to customer data, raising privacy concerns if not handled responsibly.

Overreliance on Predictions: Relying too heavily on cancellation predictions might lead to overbooking or other issues if the predictions are not accurate.

Ethical Considerations: Using predictive analytics raises ethical questions about fairness, especially if certain customer segments are unfairly targeted based on predictions.

Algorithmic Bias: If the predictive model is trained on biased data, it may perpetuate or even exacerbate existing biases, leading to unfair treatment of certain customers.

Unforeseen Factors: Predictive models might not account for unexpected events (e.g., natural disasters, political events) leading to inaccurate predictions, and potentially financial losses if not managed carefully.

**7. APPLICATIONS**

**Hospitality Industry**: Hotels, resorts, and vacation rentals use cancellation predictions to optimize room availability, staff scheduling, and pricing strategies. This helps in maximizing revenue and improving guest satisfaction.

**Travel and Tourism**: Airlines, trains, and bus companies utilize cancellation forecasts to manage seat inventory, adjust ticket prices dynamically, and offer promotional deals to fill cancelled seats.

**Restaurants:** Restaurants can predict reservation cancellations to manage table availability, optimize kitchen staff, and improve customer experience by minimizing wait times.

**Event Management**: Event organizers use cancellation predictions to adjust seating arrangements, catering, and other logistical arrangements based on the expected number of attendees.

**Healthcare:** Hospitals and clinics can predict patient appointment cancellations, enabling them to optimize doctor schedules, reduce patient wait times, and allocate resources efficiently.

**Car Rentals**: Rental car agencies use cancellation forecasts to manage their fleet, ensuring they have the right number of cars available at any given time, thus optimizing operational costs.

**Gyms and Fitness Centres**: Fitness centres use cancellation predictions to optimize class schedules, staff shifts, and equipment availability, ensuring a smooth experience for their members.

**Online Marketplaces**: Platforms offering services like home rentals, car-sharing, and experiences can optimize their offerings based on predicted cancellations, ensuring both providers and customers have a seamless experience.

**Education**: Educational institutions use cancellation predictions for managing class sizes, allocating classrooms, and organizing events, ensuring efficient use of resources.

Spas and Wellness Centres: Similar to gyms, spas and wellness centres use cancellation forecasts to optimize appointments, staff schedules, and inventory of products and treatments.

**8. CONCLUSION**

Cancelled reservations are a real big headache for hotels, and they are one of the reasons that a lot of hotels loose revenue and profits. As we all saw online booking websites are encouraging more and more customers to book more hotels and then decide which one they will stay, participating in the increase of the number of cancellations we all saw in the last few years. This study contributes to the literature on hotel booking cancellations by adopting the model integrating based on finite predictors and the Lasso regression with interactive feature analysis to support hotel cancellation predication. We use the proposed prediction method and conduct two computational experiments with real-world data to predict the individual hotel booking cancellation.

**9. FUTURE SCOPE**

The future scope of reservation cancellation prediction is promising, especially in industries like hospitality, transportation, and event management. Here are some potential areas of growth and development:

Improved Accuracy: Advancements in machine learning and data analytics can lead to more accurate cancellation predictions. This can help businesses optimize their resources and reduce revenue loss due to cancellations.

Real-time Predictions: Integrating real-time data and AI algorithms can enable businesses to predict cancellations on the fly, allowing them to take proactive actions to mitigate losses.

Customized Solutions: Tailoring cancellation prediction models to specific industries and businesses can enhance their effectiveness. For example, hotels, airlines, and event organizers may require unique models.

Customer Experience Enhancement: Businesses can use cancellation predictions to provide better customer service by offering alternatives or incentives to prevent cancellations.

Revenue Management: By accurately predicting cancellations, revenue managers can adjust pricing strategies, overbooking policies, and inventory management to optimize revenue.

**10. BIBLIOGRAPHY**

Hotel booking through online travel agency: optimal Stackelberg strategies under customer-centric payment service

Predictive analytics for customer repurchase: interdisciplinary integration of buy till you die modelling and machine learning

Booking horizon forecasting with dynamic updating: a case study of hotel reservation data.

**11. APPENDIX**

**SOURCE CODE**

# Import necessary libraries

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import LabelEncoder

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import accuracy\_score, confusion\_matrix

# Load your dataset

data = pd.read\_csv('reservation\_data.csv') # Replace with your dataset file

# Data preprocessing

# Assuming you have features like 'booking\_date', 'guests', 'room\_type', etc.

# You'll need to transform categorical data (like 'room\_type') into numerical using Label Encoding

le = LabelEncoder()

data['room\_type\_encoded'] = le.fit\_transform(data['room\_type'])

# Define the features (X) and target variable (y)

X = data[['booking\_date', 'guests', 'room\_type\_encoded']]

y = data['cancellation'] # 'cancellation' is the target variable

# Split the data into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Create and train the logistic regression model

model = LogisticRegression()

model.fit(X\_train, y\_train)

# Make predictions on the test set

y\_pred = model.predict(X\_test)

# Evaluate the model

accuracy = accuracy\_score(y\_test, y\_pred)

confusion = confusion\_matrix(y\_test, y\_pred)

print(f'Accuracy: {accuracy}')

print(f'Confusion Matrix:\n{confusion}')