DSC 630 Milestone 2: Data Selection and Project Proposal

Chitramoy Mukherjee

Predictive analysis on Hotel Booking Cancellation using hotel_booking.csv dataset

Predictive analysis in hotel booking cancellation involves using historical data and statistical techniques to forecast the likelihood of a guest canceling their reservation before their scheduled arrival date. In this project, we aim to build a predictive model to determine whether a hotel booking would be canceled, which is crucial for hotels as cancellations affect revenue and operational planning. Dataset contains many features related to booking, such as lead time, deposit type, and special requests, which adds to the model's complexity. This prediction model will help the business to determine the probability of cancellation of a booking and create a backup plan to overcome any loss. From a customer perspective we can analyze what will be the best time to book a hotel and get the best pricing for it.

Dataset information and steps:

Hotel_booking.csv dataset contains 2015-2017 timeframe data for City hotel and Resort hotel booking. Below are the variables and its description from hotel_booking.csv dataset.

Index	Variable	Description
1	**hotel**	Type of hotel (Resort Hotel, City Hotel)
2	**is_canceled**	Reservation cancellation status (0 = not canceled, 1 = canceled)
3	**lead_time**	Number of days between booking and arrival
4	**arrival_date_year**	Year of arrival
5	**arrival_date_month**	Month of arrival
6	**arrival_date_week_number**	Week number of the year for arrival

7	**arrival_date_day_of_month**	Day of the month of arrival
8	**stays_in_weekend_nights**	Number of weekend nights (Saturday and Sunday) the guest stayed or booked
9	**stays_in_week_nights**	Number of weeknights the guest stayed or booked
10	**adults**	Number of adults
11	**children**	Number of children
12	**babies**	Number of babies
13	**meal**	Type of meal booked (BB, FB, HB, SC, Undefined)
14	**country**	Country of origin of the guest
15	**market_segment**	Market segment designation
16	**distribution_channel**	Booking distribution channel
17	**is_repeated_guest**	If the guest is a repeat customer (0 = not repeated, 1 = repeated)
18	**previous_cancellations**	Number of previous bookings that were canceled by the customer
19	**previous_bookings_not_canceled**	Number of previous bookings that were not canceled by the customer
20	**reserved_room_type**	Type of reserved room
21	**assigned_room_type**	Type of assigned room
22	**booking_changes**	Number of changes made to the booking
23	**deposit_type**	Type of deposit made (No Deposit, Refundable, Non Refund)
24	**agent**	ID of the travel agent responsible for the booking
25	**company**	ID of the company responsible for the booking
26	**days_in_waiting_list**	Number of days the booking was in the waiting list
27	**customer_type**	Type of customer (Transient, Contract, Transient-Party, Group)
28	**adr**	Average Daily Rate
29	**required_car_parking_spaces**	Number of car parking spaces required
30	**total_of_special_requests**	Number of special requests made
31	**reservation_status**	Last reservation status (Check-Out, Canceled, No-Show)
32	**reservation_status_date**	Date of the last reservation status
33	**name**	Guest's name
34	**email**	Guest's email address
35	**phone-number**	Guest's phone number
36	**credit_card**	Last four digits of the guest's credit card

Below are the key steps involved for project execution,

Historical Data: Historical booking data serves as the foundation for predictive analysis. It includes information on past reservations, cancellations, guest demographics, booking channels, and other relevant variables. Data needs to be cleaned up and handled and loaded for further processing.

Feature Selection: Identifying relevant features or variables that influence cancellation behavior is crucial. These may include lead time, booking channel, seasonality, room type, price, guest demographics, and external factors such as events or holidays.

Model Development: Predictive models, such as logistic regression, decision trees, random forests, or neural networks, are trained on historical data to predict the likelihood of cancellation for future bookings. These models learn from patterns and relationships in the data to make predictions.

Model Evaluation: Models are evaluated using metrics such as accuracy, precision, recall, or area under the ROC curve (AUC) to assess their predictive performance. Cross-validation techniques help validate the model's generalizability.

Implementation and Deployment: Once a predictive model is developed and validated, it can be deployed into hotel management systems to generate real-time predictions for upcoming bookings. These predictions inform decision-making processes related to pricing, inventory management, and resource allocation.

Significance:

Predictive analysis helps hotels optimize pricing strategies by adjusting room rates dynamically based on predicted cancellation probabilities and demand fluctuations. Anticipating cancellations allows hotels to better manage inventory, allocate resources efficiently, and minimize the impact on operations. By proactively managing cancellations, hotels can minimize overbooking situations and ensure a smoother guest experience, ultimately leading to higher guest satisfaction and loyalty. Predictive insights enable hotels to make informed decisions

regarding marketing campaigns, promotions, and capacity planning, thereby maximizing revenue potential and competitiveness in the market.

Types of Models:

For this project, I plan to utilize several machine learning models to predict the chances of cancellation of hotel booking. Will implement and tune classification models including Decision Trees, Random Forest, and XGBoost. Will Emphasize achieving high F1-score for class 1, ensuring comprehensive identification of booking cancellations.

Evaluation of Results:

Will select evaluation metrics suitable for the binary classification problem of cancellation prediction. Common metrics includes below,

- Accuracy: Measures overall correctness of predictions.
- Precision: Indicates the proportion of predicted cancellations.
- Recall: Measures the proportion of actual cancellations that are correctly predicted.
- F1 Score: Harmonic means of precision and recall, useful for imbalanced datasets.
- Area Under the Receiver Operating Characteristic Curve (AUC-ROC): Measures the model's ability to discriminate between positive and negative classes.

Learning Objectives:

Through this project, I aim to gain a deeper understanding of predictive analytics techniques and their application Identify trends and factors influencing booking cancellations to anticipate future cancellations more accurately. Gain insights into optimizing revenue management and resource allocation strategies based on predicted cancellation probabilities. Understand guest behavior and

preferences regarding cancellations to enhance guest satisfaction and improve overall hotel operations.

Execution and Management of Project

Project execution Plan :

Week 3-5

- Data understanding and Exploratory data analysis.
- Data preparation and feature engineering.
- Finalize model scoring metrices.
- Exhaustive Model training and model evaluation and selection.
- Start working on final documentation.

Week 6-9

- Fine tune the models and finish any left-over work.
- Review comments from peer and instructor.
- Prepare executive summary and presentation stating the intermediate outcomes and results.

Week 10-12

- Implement peer reviews and instructor’s comments.
- Prepare the final presentation.
- Work on audio recording with slide show.
- Final submission.

Risks and Ethical Implications:

One of the major risks for utilizing guest data for predictive analysis raises privacy risks, necessitating strict adherence to data protection regulations and ensuring secure handling of sensitive information to prevent unauthorized access or misuse. There's a risk of algorithmic bias leading to unfair treatment of certain groups, such as discriminating against guests from specific demographics or regions. Ethical considerations demand the identification and mitigation of biases to ensure fair and equitable predictions. Lack of transparency in the predictive modeling process can erode trust and lead to unintended consequences. Ensuring transparency, accountability, and clear communication about the model's limitations and implications is essential to maintain ethical standards and stakeholder trust.

Contingency Plan:

If the original project plan does not work out as expected, I will reassess the data sources and modeling techniques to identify alternative approaches. This may involve exploring additional datasets, adjusting preprocessing steps, or experimenting with different machine learning algorithms. Regular communication with project advisors and peers will help in troubleshooting and adapting to any challenges encountered during the project. If I must change the topic if the results are not expected, I will work on the breast cancer prediction modeling. I have checked the dataset on that which I can use for this project.

Additional Considerations:

In addition to developing predictive models, I plan to explore feature importance to understand which variables contribute most to readmission risk. Furthermore, I will visualize model predictions and insights to facilitate interpretation and decision-making by healthcare

professionals. Continuous refinement of the predictive models based on real-world feedback and new data will be essential for improving their accuracy and applicability in clinical settings.

References:

Dataset hotel_booking.csv is sourced from https://www.kaggle.com.