Building a prediction model for hotel reservation cancellations based on historical data

Authors: Jan Erik Jakstein, Mattias Rahnu
(Group E7)
GitHub Repository

Business understanding

Background

A significant number of hotel reservations are called-off due to cancellations or no-shows. With the option to make a reservation for a hotel via online, the booking possibilities and customers' behaviour has changed dramatically. The reasons for a cancellation vary, typical causes are change of plans or scheduling conflicts or something alike. Oftentimes the customer can cancel their reservation free of charge or at a low cost. This makes it easier for the customer to cancel their booking, but causes a problem for the hotel - every canceled reservation is a possible loss in revenue that hotels have to deal with.

Business goals

This project could benefit every hotel that needs to reduce or predict their reservation cancellations. The business goal is to mitigate the revenue loss caused by reservation cancellations for a hotel by predicting whether or not the customer is going to honor their reservation. By documenting the steps and ideas of this project, hotels can reproduce the results and tailor the steps to their specific needs and constraints. Enabling them to account for possible

Business success criteria

A successful model would be able to provide hotels with enough information to cause the revenue to increase in a statistically significant way. Meaning that the likelihood of change in revenue being caused by the use of prediction model would be significant (p-value = 0.05).
Potential Revenue Mitigation: Demonstrate potential to reduce revenue loss by a minimum of 5% in historical or simulated scenarios.

Inventory of resources

- Kaggle dataset consisting of historic reservation data of two hotels.
- Anaconda and Jupyter Notebook software.
- Common Python data science libraries for model creation and optimization.
- A team of 2 data miners.
- Laptops to be able to use the software.
- Technical support and guidance from the university.

Requirements, assumptions, and constraints

Deadline: Monday, 8th of December. Finished work is considered acceptable when set goals are met, all necessary documentation points are covered and an explanatory poster has been made.

Risks and contingencies

Causes that potentially could delay the completion:

- Poor time management - Suitable timetable for different tasks and a project plan.
- Loss of work - Version control.

Terminology

Classification model - A supervised machine learning algorithm that assigns data points categories or classes.

Supervised learning - A machine learning technique that uses labeled data to train algorithms to make predictions or classifications.

AUC score - Area Under the Curve score is a performance metric for classification models, measuring how well a model can distinguish between two classes, with values mainly ranging from 0.5 to 1.0.

Costs and benefits

No cost or benefit, irrelevant for this project.

Data-mining goals

This project aims to deliver steps for creating a model with supervised learning that could predict if a customer is going to cancel their reservation or honor it. With the help of the created classification model, the project aims to give insight on how to discover which factors affect the searched outcome the most. Finally, some insights on how to mitigate the cancellation rate will be given. All of this will be delivered in a comprehensive report explaining the steps, with reasons and ideas behind them, and steps for creating this type of model. The processed sample dataset, model and a conclusive poster with ideas and information of this project will also be delivered.

Data mining success criteria

- Provide at least three distinct, feasible intervention strategies tied to the model's predicted risk levels.
- The predictive model must achieve a minimum AUC score of 0.85 on test data.
- True Positive Rate**:** The model must correctly flag a minimum of 75% of actual cancellations.

- Feature Quantification: Identify and rank the top 5 most influential factors driving cancellation probability.

Data Understanding

Data gathering

This project uses a dataset consisting of historical reservation data of two Portuguese hotels from the year 2017-2018. As this project's aim is to create a roadmap for other hotels to use on their own internal data, this dataset is used as a sample. The dataset is publicly available and accessible on kaggle. The data is in csv format.

Necessary data types:

- Outcome - cancellation or no cancellation.
- Selected time and duration for the stay.
- Some info about the customer - how many people are staying, has this guest stayed in the hotel before etc.
- Some info about the chosen reservation - for example price of the reserved room, number of special requests, required car parking space etc.

In order to address the data mining goals, the hotel reservation dataset must have a minimum time-range of one year, since customers' decisions may depend on at what time of the year the reservation was made.

Since the kaggle dataset is in csv format and can be used with Python in Jupyter Notebook and the dataset loads without setbacks, it is possible to continue with describing the data.

Data describing

The dataset used in this project consists of 36275 rows and 19 columns. The field names are the following:

- no_of_adults and no_of_children per room;
- no_of_weekend_nights and no_of_week_nights that are in the duration of a reservation;
- lead_time, which is the time elapsed between the Booking Date and the Arrival Date;
- arrival_year, arrival_month and arrival_date;

- no_of_previous_cancellations and no_of_previous_bookings_not_canceled;
- avg_price_per_room;
- no_of_special_requests;

These were all numeric fields. The following are all categorical:

- Booking_ID for different reservations, in a format such as INN00000;
- type_of_meal_plan - either no meal is selected or a meal plan 1-3 that the sample hotel is offering is selected.
- required_car_parking_space, where 1 is yes and 0 is no;
- repeated_guest, where 1 is yes and 0 is no.
- room_type_reserved, this hotel dataset has rooms from 1 to 7.
- market_segment_type, meaning if the booking was made online, offline - for example with a direct call to the hotel or maybe a walk-in without a premade reservation, corporate, meaning guests who travelled for business and whose business has an agreement with potential benefits with the hotel, complementary - room was deeply discounted, and also aviation - reservations made for airline crews.
- booking_status - either canceled or not canceled.

The dataset from kaggle falls into the set requirements, all fields are expected and relevant and there is no missing data, therefore the dataset is suitable for the next steps.

Data exploring

The main tactic chosen for exploring reservation data is to view how different attributes affect cancellation rate. Another factor to consider would be to check if there are attribute values with very few instances and if yes, then find a reason why that might be. For example, checking the time period for these instances might reveal that this value, which can be a room or meal plan or something alike, is a new offer from the hotel. Therefore, feature engineering a datetime will be useful here.

When exploring data, it can be seen that when a complementary booking has been made, there is not a single instance where a booking has been canceled. This is useful to know, as it might suggest money plays a role in whether a cancellation takes place or not, but these instances offer no other value for predictions. In total there are 391 instances of complementary reservations, which is 0.01% of all 36275 reservations in the dataset.

When viewing meal plans offered by the hotel, it can be seen that meal plan 3 is chosen significantly less. This could indicate that it is a special type of meal plan. The same can be seen with Room type 3, which has very few bookings.

As there are some non-binary categorical fields in this dataset, then these should be converted to binary fields to be able to train a model effectively.

- no_of_adults - range(0, 4) 2 most likely, 0 adult reservations could be errors.
- no_of_children - range(0, 10) left skewed with average of 0.1, there aren't many families staying at the hotel, potential errors in data: only two reservations with 9 children and one with 10.
- no_of_weekend_nights - range(0, 7) left skewed with mode being 0, two week stays are rare.
- no_of_week_nights - range(0, 17) mostly one week stays with a 3 day average, more than 5 days is rare.
- required_car_parking_space - 3% of reservations need a parking space.
- lead_time - range(0, 443) left skewed, exponential fall off.

No significant problems with the data were found.

Data verifying

Since there are no severe data quality issues and only 0.01% of instances are of no value for predictions, then the data is sufficient and can be used in the next phase - data preparation.

Project plan

Data preparation

Data selection, data cleaning, one-hot-encoding and deriving new necessary fields, recurring tasks throughout the preparation stage. Documentation of the process to make it reproducible and manageable.

Both team members, each 6h.

Repository and code management

Creating a codebase that is easy to understand and navigate. Steps in the project should be laid out in a logical order and work without having to know everything.

Both team members, each 4h.

Modeling

Test and choose the best modeling technique.
Document chosen modeling techniques and modeling assumptions.
Design tests for models (recurring) - Creating test and validation datasets for models to avoid overfitting when searching for best techniques.
Output models for testing and evaluation.

Both team members, each 8h.

Evaluation of results

Using success criteria evaluate models abilities.
Find patterns and parameters with strong correlation and analyze them.
Outline most important findings.

Both team members, each 6h.

Presentation

Creating and presenting a poster with necessary requirements at the data science project showcase session. Poster presentation should include a little introductory speech for people interested.

Both team members, each 6h.

Tools used

- Python
- Jupyter notebook
- Visual Studio Code
- Canva
- Google Docs