A logo for college computing

Description automatically generated

# TrackML: Elevating Revenue Performance Through Machine Learning and Deep Learning Techniques.

|  |  |
| --- | --- |
| *Student Name:* ***Giorgio Zoppi*** |  |
| *Student Number:* ***sba2301*** |  |
| *Module Name:* ***Strategic Thinking (HDip in Data Analytics - Feb 2024 - HCI cohort)*** |  |
| *Assignment Title: CA2-*  ***Capstone Project Proposal*** |  |
| *Assessment Due Date: 19****th May 2024*** |  |
| *Date of Submission 19****th May 2024*** |  |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Contents

[TrackML: Elevating Revenue Performance Through Machine Learning and Deep Learning Techniques. 1](#_Toc162431480)

[*Module Name:* ***Strategic Thinking (HDip in Data Analytics - Feb 2024 - HCI cohort)*** 1](#_Toc162431481)

[Introduction 3](#_Toc162431482)

[Understanding hotel room demand over time. 3](#_Toc162431483)

[Prediction of booking cancellations. 3](#_Toc162431484)

[Online hotel reputation. 3](#_Toc162431485)

[Problem Domain and Objectives 3](#_Toc162431486)

[Demand forecasting. 3](#_Toc162431487)

[Booking Cancellations. 4](#_Toc162431488)

[Online Reputation. 4](#_Toc162431489)

[HotelRank Score. 4](#_Toc162431490)

[Scope. 4](#_Toc162431491)

[Data Sources. 5](#_Toc162431492)

[Ethical Considerations. 6](#_Toc162431493)

[References 6](#_Toc162431494)

# Introduction

Revenue management (Ivanov, 2014) is a very important to make profits in the hotel industry, three main factors play an important role to get it right:

* Hotel room demand over time.
* Prediction of booking cancellations.
* Online hotel reputation.

### Understanding hotel room demand over time.

Accurate demand forecasts enable hotels and revenue managers to adjust prices dynamically, thus maximizing revenue potential. However, factors such as location, cultural events, weather, seasonal patterns and more, significantly impact on optimal hotel room forecasting (Apostolos, 2021).

### Prediction of booking cancellations.

Booking cancellations (Nuno Antonio, 2019) are an issue for the hotel revenue manager because they make harder to predict the number of booked rooms. A common pattern to address this issue is trying to achieve the overbooking by lowering the price but when overbooking is real, it is a problem because it damages hotel reputation and revenue. People might find themselves without a room and complain online. So, from a revenue manager perspective, it is important having cancellations predictions.

### Online hotel reputation.

Using the framework developed in (Diana-Jens & Rodríguez Ruibal, 2015) , we define what means online hotel reputation. Online reputation is *“the result of what clients, former clients, future clients, employees, etc. say, write and communicate to another anywhere in the internet social media based on their perceptions and experience in any moment of their relationship, direct or indirect, with the brand”*. As demonstrated in a Chinese study on ten luxury hotels (Wang, et al., 2023), negative online reviews can lead to a low booking rate that takes months to recover, so online reputation is an important factor to consider.

Our focus is therefore to create a hotel ranking algorithm called **HotelRank** that considers all these three factors in a weighted manner.This research project will be executed in strict collaboration with Blastness Spa(Blasteness.com, 2023)**,** the #1 provider in Italy for luxury hotels with a portfolio of over nine hundred hotels.

# Problem Domain and Objectives

Our main goal is to increase the profit for our customers that are mainly hotel revenue managers providing a way to compare with competitors. To achieve our main goal, we want to put our focus on creating models for demand forecasting, booking cancellation and online reputation and then combining them to create a global hotel score to understand hotel performance.

## Demand forecasting.

Our demand challenge here is to validate our dataset with data about weather and events obtained using a public API (i.e. weather.com and predicthq.com), cleaning the data, selecting the model, evaluate his performance and predict results. Our hypothesis to validate here is that demand depends on historical data and external facts (i.e. weather, events and so on).

## Booking Cancellations.

Using the same process as we plan to use for the demand, here we want to understand how cancellations affect the demand.

## Online Reputation.

Here our objective is to collect TripAdvisor data for the hotels in our dataset and see how the reviews and comments, providing our own reputation score. The hypothesis to validate here is that higher is the rank, higher is the revenue as some studies state. (Diana-Jens & Rodríguez Ruibal, 2015)

## HotelRank Score.

Once we’ve created the models, we can define an iterative process to compute **HotelRank** ranking score for each hotel. Let’s define **HotelRank** as weighted combination linear between those factors:

+ *w4 \* OptionalScore*

After model training, the correct weights **w1, w2, w3** are set and **w4** to zero and reserved for future uses.

Once **HotelRank** is defined our goal becomes to make models predictions and hotel ranking accessible to our customers in Azure.

# Scope.

Project management CRISP-DM methodology will be used most of the project except during deploy where we switch to Scrumban (Alliance, 2017). An important point in CRISP-DM is that it is meant to be an iterative process since the construction of a model requires several cycles. Over the arc of two semesters the scope of the project will try to answer the following questions:

* Which is most accurate model for demand forecasting?
* Can external facts impact the demand?
* How does cancellations affect revenue performance?
* How does hotel online reputation affect revenue performance?
* Can we determine how good we are respect our competitors?

In the table below we summarize key milestones. The deadlines might have some minor deviations due to the project complexity, so we’ve kept one month buffer at the end.

|  |  |  |
| --- | --- | --- |
| **Phase** | **Objectives** | **Milestone Deadline** |
| **Hotel Domain Knowledge Research** | Domain Analysis. Understand how ML techniques are used. Understand how Deep Learning are used in the domain. Report about domain knowledge. | **20th April 2024** |
| **Data Collection** | The dataset consists of the bookings of eight Italian luxury hotels in a two-year period. Data Collection from Weather.com. Data Collection from PredictHP.com. Data Collection from TripAdvisor. All data will be in a data lake to be able to have further processing. | **21st May 2024** |
| **Data Exploration.** | Explore common proprieties in the datasets. First cleaning the data, remove all Italian references. Visualization of the datasets. Understanding data patterns. | **4th June 2024** |
| **Data Quality Checks.** | Check the quality of data. | **15th June 2024** |
| **Feature Engineering** | Create a merged datasets to include weather and events. Cross reference hotel and user reviews. Select/Add/Remove features. | **15th July 2024** |
| **Cleaning Data** | Handling missing values. Handling Duplicates. Assure Data Consistency. | **1st August 2024** |
| **Integrate Data in Iceberg Tables** | Once the data is clean format in a query able data source to facilitate training. | **1st August 2024** |
| **Modelling: Demand and cancellation forecast models using ML.** | Creating models, training and evaluating their performance using an iterative approach using XGBoost, Regression, RNN, LSTM and Prophet algorithm. Detect overfitting. K-Fold Cross validation to determine which model performs better. | **1st September 2024** |
| **Hyperparameters tuning.** | Tune ML and deep learning models hyperparameters and test visualization. | **15th October 2024** |
| **Compute Reputation Score** | Top reviews analysis and score computation between the hotel in the dataset using classification algorithms. | **15th November 2024** |
| **Compute HotelRank.** | Select the models and weight for HotelRank and perform the computation on the dataset. Classification of the hotels using HotelRank. | **15th December 2024** |
| **Data Engineering Automation.** | Deploy the selected models in the cloud.  Automate all data flow from ingestion to model training. Provide model access to hotel revenue managers via REST API. | **20th January 2025** |
| **Reporting and Project Close.** | Project report with summary of the results indicating the chosen models and the process.  Project review: Retrospective document to indicate what went well and which are the areas of improvement. | **1st February 2025.** |

# Data Sources.

The data has been provided and released by Blastness.com on Creative Common License in the GitHub repository <https://github.com/CCT-Dublin/capstone-project-feb-2024-pt-giorgiozoppi> . The folder **hoteldataset** that contains bookings on eight luxury hotels, during a period of two years at least.

Blasteness.com (Blasteness.com, 2023) has also provided the latitude and longitude of each hotel to cross reference with TripAdvisor, PredictHQ.com and Weather.com API but this will not be disclosed for keeping the hotel name private. We are going to use TripAdvisor API for collecting the reviews and store them in a data storage to be processed later and create a dataset to use for the online reputation score. Weather.com API and PredictHP.com API data will be used to enrich the bookings dataset.

# Ethical Considerations.

Hotel revenue management can have a positive influence on the society. A socially responsible hotel management could use this study promote a fair or a festival in the moment of maximum demand, creating a nice environment for locals and tourists or reduce their environmental impact when the demand is at its minimum (i.e. waste less water or electricity, less trash to recycle and so on).

The data on this study has been anonymized to remove all hotel and customer confidential data. Any data and study taken in consideration for learning the field will be cited in the report under Harvard guidelines and avoid plagiarism. All the model developed will be put in production on Azure, also Blastness(Blasteness.com, 2023) will have the right to cite this work as joint effort.

The repository data has been release with Creative Common License at <https://github.com/CCT-Dublin/capstone-project-feb-2024-pt-giorgiozoppi> where you can find this report as well. An authorization mail has been provided.

Immagine che contiene testo, algebra, schermata

Descrizione generata automaticamente

# Introduction to Data Exploration.

As stated in the milestones, the data exploration phase is due the first week of June. At this moment we’ve not yet collected the data from external sources such as TripAdvisor, PredictHQ.com and Weather.com but we can focus directly on the current Blastness dataset and booking forecasts to compute the ***DemandScore*** part of the Hotel Rank. The scope is to choose the best algorithm that is able to compute the ***DemandScore***.

# Data Understanding.

Initially we must restrict the features of the dataset since the CSV file come directly from SQLServer. Once we’ve done this and done some initial explorative analysis. We’ve discovered a lot of fields that were not needed. At this point we asked ourselves about the distribution of arrival and the dataset.

Immagine che contiene testo, diagramma, Diagramma, schermata

Descrizione generata automaticamente

The arrival intervals follow a normal distribution this allow to predict arrivals easily. In normal distribution arrival times are centred around the mean with the spread determined by the standard deviation. This means that it easy to predict hotel occupancy. For example, if the average number of daily arrivals is 50 with standard deviation of 10, most days will see between 40 and 60 arrivals. We’ve also seen during data exploration several columns that were not correct and fixed the data. During this process we’ve detected several NaN columns that we later correct in the data preparation.,

# Data Preparation.

In this phase we had to clean the data remove NaN, distinguish categorical data and numerical to see the correlation matrix between the data. We selected the following rows:

* **Code:** Booking code.
* **Status:** Booking Status.
* **Arrival:** Date of the arrival.
* **Departure:** Date of the departure.
* **Nights:** Number of the night booked
* **BookingChannel:** Where is booked (Booking.com, Hotel WebSite, etc.)
* **Total:**  Total Price.
* **PurchaseDate:** Date of the purchase.
* **BookingDevice:** Device from has been booked.
* **LastModified**: Date of the last modification.
* **HotelId:** Id of the Hotel.

From these initial features we had to add the city provided separately, computed the season.

Once we had this, we had to merge each dataset in one bigger, so at the end we’re considering ten luxury hotels in Italy in the last two years.

We want to see the correlation matrix to see the correlation matrix to see which features we could use in the future when we will compute the other scores. Now we will focus at first on the demand one, so basically arrival forecasting.

To compute the correlation matrix we had to reduce the feature and perform one shot encoding on categorical features. We have to reduce the number of features, in particular we’ve dropped: ‘Code’, ‘Arrival’, ‘Departure’, ‘PurchaseDate’,’LastModified’ because we replaced them with their timestamp or were redundant.

We’ve added numerical timestamp, city of the hotel provided from Blastness/Fox technology. We also had to remove null values and replace with unknown where opportune.

# Modelling.

We’ve identified two models: Prophet and SARIMA. Our objective was to compute DemandScore on our dataset.

## Prophet.

Prophet, an open-source tool developed by Facebook's Core Data Science team, is designed for large-scale, automated time series forecasting. Prophet requires minimal data preprocessing due to its robustness against outliers, missing data, and significant changes in historical trends. It also offers tuneable, human-interpretable parameters based on business insights and allows manual addition of changepoints (sudden trend shifts), which is particularly useful when known business events are likely to impact trends.

Prophet is an additive model that decomposes a time series into three components: trend, seasonality, and holidays. Trend function g(t) estimates the non-periodic changes in the series. Seasonality s(t) represents any periodic changes, which could be daily, weekly, yearly, or any custom periods. Holiday h(t) represents the effects of holidays that happen irregularly.  y(t) = g(t) + s(t) + h(t) + error.

## SARIMAX.

SARIMAX (Seasonal Auto-Regressive Integrated Moving Average with Exogenous variables) is a traditional statistical time series model based on the ARIMA model. It uses the target variable's past values (auto-regressive terms) and past errors (moving average terms) to forecast future values. For non-stationary time series, SARIMAX handles shifts by using differences between consecutive time steps. The model incorporates seasonal auto-regressive and moving-average terms, analysing values and errors at multiples of the seasonal period 𝑠s.

We have run prediction on both cases, and we’ve seen that in this dataset SARIMAX work better. The values of Mean Average Error, Mean Square Error and Root Mean Square error are better.

* **Prophet** MAE: 97.22875881839182, MSE: 13989.777902305872, RMSE: 118.27839152738709
* **SARIMA** MAE: 103.55959060112583, MSE: 13553.634110175268, RMSE: 116.42007606154219

Those are the two models used in literature for this problem. We’ve cleaned the data and discovered on this dataset that SARIMAX perform better, and we will use it for computing **DemandScore**.

To identify the performance, we’ve done predictions for the next 52 weeks and computed MSE, MSE and RMSE. A better approach to experiment in future is to use the same algorithm on more years and split testing and prediction parts.

# Conclusion.

We’re able to compute Demand Score for HotelRank, but still some gaps are present:

* Collection with external events.
* More data to decrease MSE.

Those will be the main objectives of the next development that we’re aiming for. We’re a bit far behind the schedule but we’ll push forward.

In this submission we’ve see which algorithm is to use for demand forecasting based on Arrival Date and we’ll use for Hotel Rank. We’ve also seen the complex structure of the data provided by Blastness, cleaned up and noticed its own peculiarities that we will use in future.

# References

Alliance, S., 2017. *What is Scrumban?.* [Online]   
Available at: https://www.agilealliance.org/scrumban  
[Accessed 03 03 2024].

Apostolos, A., 2021. Modeling and Forecasting Daily Hotel Demand: A Comparison Based on SARIMAX, Neural Networks, and GARCH Models. *Forecasting*, 21 August, p. 16.

Blasteness.com, 2023. *Blastness - Growing your business.* [Online]   
Available at: https://www.blastness.com/en/index  
[Accessed 26 03 2024].

Diana-Jens, P. & Rodríguez Ruibal, A., 2015. La reputación online y su impacto en la política de precios de los hoteles. *Cuaderno de Turismo*, 15 07, p. 129–155.

Ivanov, S., 2014. *Hotel revenue management: From theory to practice.* 1 ed. Varna, Bulgaria: Zangador Ltd..

Nuno , A., De Almeida , A. & Nunes, L., 2019. Hotel booking demand datasets. *Data in Brief,* 22(February), pp. 41-49.

Nuno Antonio, A. D. A. L. N., 2019. Big data in hotel revenue management: Exploring cancellation drivers to gain insights into booking cancellation behavior. *Cornell Hospitality Quarterly,* 60(4), pp. 298-319.

Wang, Z., Lin, X. & Li, H., 2023. Impact of reputation on hospitality profitability: impact of service failure online exposure on revenue performance – evidence from the hotel industry in China. *Tourism Review*, 17 3, pp. 1387-1413.