

Charles Malon Silva Rocha - 2021376

Lucas dos Santos Barbosa - 2021337

**Problem Solving for Industry**

**Capstone Group Project**

**Dublin**

**2025**

**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a Word doc for students to include with every submission.*

|  |  |
| --- | --- |
| **Module Title:** | Problem Solving for Industry |
| **Assessment Title:** | Capstone Group Project |
| **Lecturers Names:** | Muhammad Iqbal, Ken Healy |
| **Student Full Name:** | Charles Malon Silva Rocha |
| **Student Number:** | 2021376  2021337 |
| **Assessment Due Date:** | 18th May 2025 |
| **Date of Submission:** | 18th May 2025 |

Below you can access the progress of this assignment.

<https://github.com/CCT-College-Dublin/ca2-90-CharlesMalonRocha>

**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

[Abstract 5](#_Toc3829)

[Project Objectives 6](#_Toc19489)

[Roles and Responsibilities 7](#_Toc20937)

[Stage 1 - Business Understanding 8](#_Toc27386)

[1.1 Project Objectives 8](#_Toc15320)

[1.2 Stakeholders 8](#_Toc25369)

[1.3 Deliverables 9](#_Toc4664)

[1.4 Impact to Target Operating Model 9](#_Toc32484)

[1.5 Communication Approach 9](#_Toc20663)

[1.6 Responsibilities 9](#_Toc16876)

[1.7 Scheduling 10](#_Toc28594)

[1.8 Technologies 10](#_Toc484)

[Our choice 11](#_Toc22962)

[1.9 Legal and Ethical Issues 11](#_Toc17479)

[1.10 Data Collection 11](#_Toc5884)

[Stage 2 - Data Understanding 12](#_Toc28214)

[Stage 3 - Data Preparation 13](#_Toc9832)

[4.1 Overview 14](#_Toc95)

[4.2 Alternative approach 14](#_Toc30277)

[4.3 SVD Algorithm 14](#_Toc9952)

[4.4 Sklearn Algorithms 14](#_Toc17687)

[4.5 14](#_Toc25992)

[4.6 14](#_Toc12647)

[Stage 5 - Evaluation 15](#_Toc10982)

[5.1 Training final model. 15](#_Toc14596)

[5.2 Testing final model 15](#_Toc10500)

[5.3 Final model accuracy in context to the final system 15](#_Toc11114)

[5.4 Overall Evaluation 15](#_Toc2238)

[Stage 6 - Deployment (or Simulation) 16](#_Toc28702)

[Conclusion 17](#_Toc29976)

[References 18](#_Toc18761)

Abstract

This project is focused on optimizing hotel overbooking using predictive modeling and simulation. The system was developed in Python, utilizing a synthetic dataset that simulates historical hotel booking data, including guest attributes, booking behavior, and no-show probabilities. The project follows the CRISP-DM methodology, with each phase thoroughly documented in a report, a poster and Jupyter Notebook.

The solution combines the strengths of machine learning and Monte Carlo simulation. We evaluate the accuracy of different models in predicting guest no-shows—such as logistic regression and random forest—and select the most effective one to integrate into our final solution. We then apply simulation techniques to test overbooking strategies and determine the optimal overbooking level that balances occupancy rates with customer satisfaction.

The result is a hybrid decision-making system that helps hotel managers increase revenue by reducing empty rooms due to no-shows, while minimizing the risk of having to turn guests away. Performance improvements and insights are visualized and clearly presented throughout the project.

Project Objectives

Roles and Responsibilities

Stage 1 - Business Understanding

*How can we predict the optimal number of overbookings without harming the customer experience?*

## 1.1 Project Objectives

The goal of this project is to develop a predictive decision support system to help hotels **optimize overbooking strategies**. Using machine learning models to predict **no-show probabilities**, and simulation methods (Monte Carlo) to evaluate risk and performance, the system aims to increase **occupancy rates** and **maximize revenue**, while minimizing the likelihood of customer dissatisfaction caused by overbooking.

By combining predictive analytics with simulation, this project proposes a **hybrid system** that reflects real-world booking uncertainty, empowering hotel managers to make more informed decisions on how many extra bookings to accept.

## 1.2 Stakeholders

· **Hotel managers and operations teams**, who need tools to balance revenue optimization with customer satisfaction.

· **Revenue management professionals**, who require data-driven insights into demand and no-show behavior.

· **Hotel chains or booking platforms**, interested in integrating predictive systems to reduce losses from empty rooms.

· **Academic institutions or research communities**, exploring applied machine learning in real-world operations.

## 1.3 Deliverables

· A machine learning model that predicts the likelihood of booking cancellations or no-shows.

· A Monte Carlo simulation module to estimate the outcomes of various overbooking levels.

· Visualizations and metrics to evaluate performance and risk.

· A prototype interface or demonstration via Jupyter Notebook.

· Supporting documentation and this report.

## 1.4 Impact to Target Operating Model

This system provides data-driven insights without interfering with existing booking infrastructure. It acts as a decision-support layer, which can be applied externally or integrated into hotel property management systems (PMS). It encourages proactive, strategic overbooking based on statistical modeling rather than intuition.

## 1.5 Communication Approach

The results and benefits of the system can be communicated through:

· Internal hotel training and presentations.

· Academic or industry publications.

· Integration into booking platforms or dashboards.

· Blogs, LinkedIn posts, or white papers for tech-savvy revenue managers.

## 1.6 Responsibilities

Our team is composed by Charles Rocha and Lucas Barbosa, both responsible for technical development, research, and reporting. Dr. Muhammad Iqbal provided technical supervision and guidance throughout the project, with additional academic support from Professor Ken Healy.

## 1.7 Scheduling

The project spans a three-month period and follows the CRISP-DM framework:

· Weeks 1–2: Business understanding and data exploration.

· Weeks 3–4: Data preparation and EDA.

· Weeks 5–6: PCA

· Weeks 7–8: Modeling and simulations.

· Weeks 9–10: Evaluation, deployment, and final reporting.

Final report and deliverables are due on May 18, 2025.

## 1.8 Technologies

Open Source Tools Used:

· Python 3.13.3

· NumPy, Pandas, and Scikit-learn for data manipulation and ML modeling.

· Matplotlib and Seaborn for visualizations.

· Jupyter Notebook for interactive development.

· SimPy and custom Monte Carlo simulations for stochastic modeling.

These open-source libraries allow full control and transparency over model development and experimentation, offering flexibility not possible in proprietary “black-box” systems.

The most fundamental libraries in Python for ML/AI projects are:

NumPy

Scikit-learn

Py-Torch

TensorFlow

Pandas

### Our choice

### 

## 1.9 Legal and Ethical Issues

### 

The dataset (hotel\_bookings.csv) is open and publicly available, and contains no personal or identifiable customer data, ensuring compliance with **GDPR** and ethical standards. In future versions of this project where live user data might be collected, considerations around data privacy, transparency, and fairness will be essential.

It’s also important to avoid **discriminatory patterns** in model predictions (e.g., if certain booking channels or countries are unfairly penalized). Ensuring ethical modeling practices and bias mitigation is a critical goal.

## 1.10 Data Collection

The data used in this project comes from a publicly available dataset on **hotel bookings**, containing over 119,000 rows with information about hotel type, booking dates, customer details, and whether the guest canceled or showed up. The data enables training of models to estimate no-show probabilities and supports simulation of hotel capacity under various overbooking levels.

Stage 2 - Data Understanding

Stage 3 - Data Preparation

Data preparation is one of the most critical steps in any machine learning project. It ensures that the dataset is clean, consistent, and structured in a way that models can interpret. For this project, we transformed the hotel\_bookings.csv dataset into a usable format for predictive modeling.Stage 4 - Modeling

### 4.1 Overview

### 4.2 Alternative approach

### 4.3 SVD Algorithm

### 4.4 Sklearn Algorithms

### 4.5

### 

### 4.6

Stage 5 - Evaluation

### 5.1 Training final model.

### 5.2 Testing final model

### 5.3 Final model accuracy in context to the final system

### 5.4 Overall Evaluation

Stage 6 - Deployment (or Simulation)

Conclusion

‌

‌

References

Iqbal, M. (2025) ‘CRISP-DM Overview’[Lecture], *Y4M8:Problem Solving for Industry.* CCT College Dublin. 17 Feb. (Iqbal, 2025)

Iqbal, M. (2025) ‘CRISP-DM stage one – business understanding’[Lecture], *Y4M8:Problem Solving for Industry.* CCT College Dublin. 24 Feb. (Iqbal, 2025)

‌

Iqbal, M. (2025) ‘CRISP-DM Case Study 1’[Lecture], *Y4M8:Problem Solving for Industry.* CCT College Dublin. 24 Feb. (Iqbal, 2025)

Iqbal, M. (2025) ‘CRISP-DM stage two - data understanding’[Lecture], *Y4M8:Problem Solving for Industry.* CCT College Dublin. 03 Mar. (Iqbal, 2025)

‌Iqbal, M. (2025) ‘CRISP-DM Case Study 2’[Lecture], *Y4M8:Problem Solving for Industry.* CCT College Dublin. 03 Mar. (Iqbal, 2025)

Iqbal, M. (2025) ‘CRISP-DM stage three – data preparation’[Lecture], *Y4M8:Problem Solving for Industry.* CCT College Dublin. 10 Mar. (Iqbal, 2025)

Iqbal, M. (2025) ‘CRISP-DM stage four – modelling’[Lecture], *Y4M8:Problem Solving for Industry.* CCT College Dublin. 24 Mar. (Iqbal, 2025)

Iqbal, M. (2025) ‘CRISP-DM stage five – evaluation’[Lecture], *Y4M8:Problem Solving for Industry.* CCT College Dublin. 31 Mar. (Iqbal, 2025)

Iqbal, M. (2025) ‘CRISP-DM stage six – deployment’[Lecture], *Y4M8:Problem Solving for Industry.* CCT College Dublin. 07 Apr. (Iqbal, 2025)

Kroese, D.P., Brereton, T., Taimre, T. and Botev, Z.I. (2014). Why the Monte Carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(6), pp.386–392. doi:https://doi.org/10.1002/wics.1314. (Kroese et al., 2014)

‌

‌

‌