**CHAPTER I – INTRODUCTION**

* 1. **Background**

Hotel reservation cancellations are on the rise in the hospitality industry due to changes in customer behavior. There can be various reasons for the cancellations like better prices in other hotels, change of plans, scheduling conflicts, etc as mentioned in (www.kaggle.com, 2023). (Antonio, de Almeida and Nunes, 2017) mentioned that to reduce the problems due to cancellations, hotels had implemented strict policies over booking cancellations which in turn has impacted the performance badly on revenue and reputation in the markets. Because of these strict policies, customers might not book a room in the same hotel in the future. Leveraging the capability of machine learning to predict whether the reservation will be canceled or not. There is one statement given by (Romero Morales and Wang, 2010), “It is hard to imagine that someone can predict whether a booking will be canceled or not with high accuracy”, but due to the advancements in data science and machine learning, its possible to predict the reservation cancellation with the help of past data obtained by hotels.

In this project, the Hotel Reservations dataset from (www.kaggle.com, 2023) is used for data analysis and experimentation with various machine learning algorithms on sampling techniques to develop a prediction model.

* 1. **Research Aim**

This project aims to identify the reasons behind the cancellations of hotel reservations and develop machine-learning algorithms on oversampled and undersampled data to predict the customer who might cancel the reservation based on the customer details provided using the machine learning with best performance.

* 1. **Research Questions**

Below are the research questions to investigate in this research work:

1) What are the columns not important for analysis and why were they dropped from the dataset?

2) What is one hot encoding and why it is crucial for analyzing this dataset?

3) What are the classification algorithms considered and which algorithm has given the best performance on comparing Under-Sampled and Over-Sampled Data?

4) Is the average price change and lead time responsible for the customer canceling a hotel reservation?

* 1. **Hypothesis**

This research would like to identify the below hypothesis:

**Hypothesis:** will over-sampled data help achieve better performance than under-sampled data on classification algorithms?

**Null Hypothesis:** There is no significant difference in the performance of classification algorithms trained on over-sampled data and under-sampled data.

**Alternative Hypothesis:** Classification algorithms trained on over-sampled data will have better performance than classification algorithms trained on under-sampled data.

* 1. **Research Objectives**

Below are the objectives of this research work:

* Understand the description of the hotel reservations dataset.
* Data Pre-processing and analyzing the data with respect to the target variable.
* Data Separation into training, validation, and testing sets.
* Understanding Logistic Regression, KNN, Decision Tree, and Random Forest.
* Understanding and implementing the sampling techniques on machine learning classification algorithms considered for this research.
* Performance comparison analysis of the algorithms on over-sampled and under-sampled data.
  1. **Project Planning**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Prediction of hotel reservation cancellation of a customer with reservation details using Machine Learning Project Planning** | | | | | |
|  |
|  |  | | | | |  |
|  |  |  |  |  |  |  |
|  | **Milestone description** | **Progress** | **Start** | **End** | **Days** |  |
| **Project Idea** | Data Source and Project Idea | **100%** | 23/02/2023 | 24/02/2023 | 2 |  |
| Data Understanding | **100%** | 23/02/2023 | 24/02/2023 | 2 |  |
| RQ's and Hypothesis Identification | **100%** | 25/02/2023 | 25/02/2023 | 1 |  |
| **Literature Background** | Hotel reservation cancellation papers | **100%** | 10/03/2023 | 12/03/2023 | 3 |  |
| Data Imbalance and Solutions | **100%** | 13/03/2023 | 14/03/2023 | 2 |  |
| RF, DCS, KNN, Logistic Regression | **100%** | 15/03/2023 | 30/03/2023 | 16 |  |
| Cross Validation Techniques | **100%** | 01/04/2023 | 02/04/2023 | 2 |  |
| **Data Cleaning** | Deal with missed values in Columns | **100%** | 03/04/2023 | 03/04/2023 | 1 |  |
| Drop any null value rows | **100%** | 03/04/2023 | 03/04/2023 | 1 |  |
| **Data Preparation and Visualization** | Understanding visualization Plots | **100%** | 15/04/2023 | 20/04/2023 | 6 |  |
| Plotting Visualizations | **100%** | 15/04/2023 | 20/04/2023 | 6 |  |
| learn code for one hot encoding | **100%** | 21/04/2023 | 22/04/2023 | 2 |  |
| One hot encode on categorical coumns | **100%** | 21/04/2023 | 22/04/2023 | 2 |  |
| **Model Building** | Split train and test | **100%** | 25/04/2023 | 25/04/2023 | 1 |  |
| Standardize the data | **100%** | 25/04/2023 | 25/04/2023 | 1 |  |
| ML algorithms on Over-Sampled data with GridSearch | **100%** | 26/04/2023 | 15/05/2023 | 19 |  |
| ML algorithms on Under-Sampled data with GridSearch | **100%** | 16/05/2023 | 30/05/2023 | 15 |  |
| Test on unseen data | **100%** | 01/06/2023 | 02/06/2023 | 2 |  |
| Performance Comparision | **100%** | 03/06/2023 | 04/06/2023 | 2 |  |
| **FPR Report** | Abstract | **100%** | 10/07/2023 | 10/07/2023 | 1 |  |
| Introduction | **100%** | 11/07/2023 | 13/07/2023 | 3 |  |
| Literature Review Search | **100%** | 14/07/2023 | 25/07/2023 | 12 |  |
| Research Methodology | **100%** | 21/07/2023 | 30/07/2023 | 10 |  |
| Results | **100%** | 01/08/2023 | 10/08/2023 | 10 |  |
| Conclusion and Future Work | **100%** | 12/08/2023 | 15/08/2023 | 4 |  |
| References | **100%** | 16/08/2023 | 16/08/2023 | 1 |  |
| Appendices | **100%** | 17/08/2023 | 18/08/2023 | 2 |  |
| Plagarism Check | **100%** | 19/08/2023 | 19/08/2023 | 1 |  |
| Final Edits and Review | **100%** | 20/08/2023 | 25/08/2023 | 6 |  |
| Final Plagarism Check | **100%** | 25/08/2023 | 25/08/2023 | 1 |  |
| Meet with Supervisor | **100%** | 26/08/2023 | 26/08/2023 | 1 |  |
| FPR Submission | **100%** | 28/08/2023 | 28/08/2023 | 1 |  |

Table 1: Project Planning

* 1. **Chapter Overview**

Below are the chapters discussed in this research project.

**Chapter I: Introduction**

This chapter discusses the background of hotel reservation cancellations and their impact and why machine learning is needed to deal with those cancellations by customers. The research aim, and research questions to be answered are mentioned, and also the hypothesis and research objectives of this project are discussed. The overall idea of this research work will be discussed in this section.

**Chapter II: Literature Review**

In this chapter, a detailed introduction to Hotel Reservation cancellation and challenges will be discussed and in the later sections, the importance of one-hot encoding, data imbalance, and its solutions will be reviewed. And later, Standardization, grid-search, and their importance for this project will be discussed followed by an overview of machine learning classification algorithms will be discussed followed by the research gap identified.

**Chapter III: Research Methodology**

The research process for this project will be discussed In this chapter starting with a detailed review of the Hotel reservations dataset and then the pre-processing techniques used and how the one-hot encoding is applied in this dataset will be discussed. Data visualizations will be interpreted and In the next sections, data separation and feature standardization will be discussed. Later, how data is over-sampled and under-sampled and applied to machine learning algorithms with the best parameters from grid-search will be discussed here. In the last, Performance metrics will be discussed followed by a summary.

**Chapter IV: Results and Discussion**

In this section, the experimentation setup used for this project and the pre-processing and data visualization results will be discussed here. Later machine learning algorithms' performance on the over-sampled data and under-sampled data will be analyzed and the results be compared to answer the research questions and hypotheses.

**Chapter V: Conclusion and Future Work**

The answers to research questions will be concluded in this chapter, which also provides a summary of the report's conclusions. It also examines if the objectives were achieved and offers some suggestions for further study.