**Table of Contents**

Abstract .............................................................................................................................. 1

Acknowledgment ................................................................................................................ 2

Declaration .......................................................................................................................... 3

List of Tables…………………………………………………………………………….. 7

List of Figures …………………………………………………………………………….. 8

CHAPTER I – INTRODUCTION ...................................................................................... 9

1.1 Background ................................................................................................................... 9

1.2 Research aim ................................................................................................................. 9

1.3 Research Questions ....................................................................................................... 9

1.4 Hypothesis …………………………………………………………………………… 10

1.5 Research Objectives ...................................................................................................... 10

1.6 Project Planning ………………………………………………………………………. 10

1.7 Chapter overview …………………………………………………………………….. 12

CHAPTER 2: LITERATURE REVIEW ............................................................................ 13

2.1 Introduction to Hotel Reservation Cancellation……..................................................... 13

2.2 What is One-Hot Encoding ……………………............................................................. 13

2.3 Data Imbalance ………………………………………………........................................14

2.3.1 How Data imbalance will impact the performance of a model ……………….. 14

2.3.2 Solutions for Data Imbalance ………………………………………………….. 15

2.3.2.1 Over Sampling ……………………………….………………………….. 15

2.3.2.2 Under Sampling …………………………………………………………. 15

2.4 Feature Scaling ……………………………………………………………………….. 16

2.4.1 What is Standardization ……………………………………………………….. 16

2.4.2 Benefits of Standardization ……………………………………………………. 16

2.5 Hyperparameter tuning with Grid Search Cross Validation………………………… 17

2.5.1 What is Grid Search Cross Validation ………………..……………………….. 17

2.5.2 Advantages and Disadvantages of Grid Search CV …………………………… 18

2.6 Overview of Machine Learning Classification .............................................................. 19

2.6.1 Logistic Regression …………………………………………………………….. 19

2.6.2 K-Nearest Neighbours .......................................................................................... 20

2.6.3 Decision Trees ….................................................................................................. 21

2.6.4 Random Forest ...................................................................................................... 22

2.6.4.1 Random Forest Feature Importance …………………………………… 23

2.7 Research Gap ................................................................................................................. 23

CHAPTER 3: RESEARCH METHODOLOGY ................................................................. 24

3.1 Introduction .................................................................................................................... 24

3.2 Hotel Reservations Dataset ……..................................................................................... 24

3.3 Data Pre-processing ........................................................................................................ 25

3.3.1 Missing Values...................................................................................................... 25

3.3.2 Removing unwanted columns.............................................................................. 25

3.4 One-Hot Encoding …………………………………………………………………….. 25

3.5 Data Visualization ……………………………………………………………………. 26

3.6 Training and Testing Data Separation ………………………………………………… 27

3.7 Feature Standardization …………………...................................................................... 27

3.8 Oversampling ………........……………………………………………………………. 28

3.8.1 Oversampling using SMOTE ………………………………………………. 28

3.8.2 Grid Search on ML Algorithms on Over-Sampled Data …………………… 28

3.8.3 Model Building with Best Params of Algorithms on Over-Sampled Data ..... 29

3.9 Undersampling ………………………………………………………………………... 29

3.9.1 Undersampling using RUS …………………………………………………. 29

3.9.2 Grid Search on ML Algorithms on Under-Sampled Data ….……………… 29

3.9.3 Model Building with Best Params of Algorithms on Under-Sampled Data ... 29

3.10 Performance Metrics …................................................................................................ 30

3.10.1 Confusion Matrix ............................................................................................... 30

3.10.2 Performance measures ….……………………………………………….......... 30

3.11 Overall Workflow ........................................................................................................ 31

3.12 Overall Summary ........................................................................................................ 31

CHAPTER 4: RESULTS AND DISCUSSION ................................................................... 32

4.1 Experimentation setup .................................................................................................... 32

4.2 Checking for missing data ............................................................................................ 32

4.3 Data Separation Results ................................................................................................. 33

4.4 Data Visualization Results ……………………………………………………………. 34

4.5 Machine Algorithms Results on Over-Sampled Data …….............................................39

4.5.1 Logistic Regression ……………………………………………………………... 40

4.5.2 K-Nearest Neighbours ………………………………………………………….. 40

4.5.3 Decision Tree ……………………….................................................................... 41

4.5.4 Random Forest ……..…………………………………………………………… 42

4.5.4.1 Feature Importance Results …………........................................................ 43

4.5.5 Combined Results of All Algorithms on Over-Sampled Data ………………… 43

4.6 Machine Algorithms Results on Under-Sampled Data ….............................................. 44

4.6.1 Logistic Regression …………………………………………………………….. 44

4.6.2 K-Nearest Neighbours …………………………………………………………. 44

4.6.3 Decision Tree ………………………………………........................................... 45

4.6.4 Random Forest …………………………………………………………………. 46

4.6.4.1 Feature Importance Results …………........................................................ 46

4.6.5 Combined Results of All Algorithms on Under-Sampled Data .………………. 47

4.7 Comparison of Algorithms Results on Over-Sampled and Under-Sampled Data ......... 47

CHAPTER 5: CONCLUSION AND FUTURE WORK ..................................................... 49

REFERENCES …................................................................................................................ 51

APPENDICIES .................................................................................................................... 54