Airline Seat Sale Prediction using Digital Click Stream Data

MD ALAUDDIN

MASTER OF SCIENCE (INFORMATION TECHNOLOGY)

MULTIMEDIA UNIVERSITY NOV 2019

Airline Seat Sale Prediction using Digital Click Stream Data

BY

MD ALAUDDIN

M.Sc. in IT (By Research), Multimedia University, Malaysia

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE (INFORMATION TECHNOLOGY)

(by Research) in the

Faculty of Computing and Informatics

MULTIMEDIA UNIVERSITY MALAYSIA

Nov 2019

©2019 Universiti Telekom Sdn. Bhd. ALL RIGHTS RESERVED.

Copyright of this thesis belongs to Universiti Telekom Sdn. Bhd. as qualified by Reg- ulation 7.2 (c) of the Multimedia University Intellectual Property and Commercialisa- tion Policy. No part of this publication may be reproduced, stored in or introduced into a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording, or otherwise), or for any purpose, without the express writ- ten permission of Universiti Telekom Sdn. Bhd. Due acknowledgement shall always be made of the use of any material contained in, or derived from, this thesis.

# DECLARATION

I hereby declare that the work has been done by myself and no portion of the work contained in this Thesis has been submitted in support of any application for any other degree or qualification on this or any other university or institution of learning.

# Your name here

**ACKNOWLEDGEMENTS**

First, I am thankful to the Almighty ALLAH for enabling me to complete this challenging task.

I would like to thank my supervisor Associate Professor Dr. Choo-Yee Ting for his supervision and experience guidance. His continuous support and encouragement throughout my candidature gave me more confidence in my work and abilities and led me to overcome all the difficulties during my study. It is my valued opportunity to learn the rigorous attitude towards study and research from him. I would also like to thank Dr. Ian Tan and for his valuable discussions, cooperation and support.

I owe every part of this work to my wife who has sacrificed a lot during my study.

To my supervisor and my family.

# ABSTRACT

Revenue Management is important for every airline business and the seat is the main product of an airline. The purpose of the revenue management is to maximize the revenue of each airline routes based on demand. This demand, however, depends on factors such as historical demand, seasonality, seat pricing based on purchase lead days, competitors pricing and customer behaviour. Prediction of passenger demand helps to estimate revenue on future flights and therefore allow the airline to generate optimal prices for the corresponding flights. Therefore, minimizing the prediction error constitutes the most crucial goal of revenue management. The accuracy of prediction models varies based on the dataset that has been used for the modelling and training purpose. Most of the previous research work ignored digital data (digital data can be defined as the data generated due to the user interaction with various digital platforms like website, mobile applications etc.) due to the lack of proper extraction and processing pipeline of this huge volume of data. Since the digital customer data provide crucial insight into customer purchasing behavior it is very important to extract useful attributes to feed the machine learning model. Furthermore, due to the various digital platform (i.e. mobile, website etc.), the structure of stored data is different. This presents a further challenge while creating an aggregated dataset by comprising different types of data sources (i.e. transactional, operational, digital). In this research, a data extraction, processing and aggregation procedure has been proposed for various digital platform (i.e. mobile, tablet, desktop, notebook etc.) used by the airline industry. To optimize the prediction accuracy, an analytic dataset has been developed by combining digital attributes and traditional operational and transactional attributes. In this research work, a Gradient Boosted Regression Tree (GBRT) model for seat sale prediction has been optimized using this dataset. The GBRT model is well suited for both classification and regression application. GBRT has been used successfully in fields like e-commerce, telecommunication, solar energy prediction etc. Therefore, GBRT has been chosen for this work. The empirical findings suggested applying GBRT on transformed dataset can predict seal sales for 30 days ahead with accuracy of 93%.

|  |  |
| --- | --- |
| **TABLE OF CONTENTS** |  |
| **[COPYRIGHT PAGE](#_bookmark0)** | **ii** |
| [**DECLARATION**](#_bookmark1) | **iii** |
| [**ACKNOWLEDGEMENTS**](#_bookmark2) | **iv** |
| [**DEDICATION**](#_bookmark3) | **v** |
| [**ABSTRACT**](#_bookmark4) | **vi** |
| [**TABLE OF CONTENTS**](#_bookmark5) | **vii** |
| [**LIST OF TABLES**](#_bookmark6) | **viii** |
| [**LIST OF FIGURES**](#_bookmark7) | **ix** |
| [**CHAPTER 1: INTRODUCTION**](#_bookmark8) | **1** |
| [1.1 Research Background](#_bookmark9) | 1 |
| [1.2 Problem Statement](#_bookmark11) | 2 |
| 1.2.1 Limited Work on Seat Sale Prediction | 2 |
| 1.2.2 Processing and Integration of Digital Data | 2 |
| 1.3 Research Motivation | 3 |
| [1.4 Research Objectives](#_bookmark13) | 4 |
| [1.5 Research Scope](#_bookmark14) | 4 |
| 1.6 Thesis Outline | 4 |
| [**CHAPTER 2: LITERATURE REVIEW**](#_bookmark17) | **5** |
| 2.1 Introduction | 5 |
| 2.2 Dynamic Pricing and Prediction | 6 |
| 2.3 Prediction methods for Airlines | 7 |
| 2.3.1 Fuzzy Prediction Method | 7 |
| 2.3.2 Artificial Neural Network (ANN) based Prediction | 7 |
| 2.3.3 GMM based Prediction | 8 |
| 2.4 Opportunity of Digital Data in Airline Industry | 11 |
| 2.5 Challenges of Digital Data | 12 |
| 2.6 Digital attributes for airline ticket sale prediction model | 12 |
| 2.7 Gradient Boosted Regression Tree | 12 |
| 2.8 Summary | 13 |
| **CHAPTER 3: DATA COLLECTION AND PROCESSING** | **14** |
| 3.1 Introduction | 14 |
| 3.2 Data Source Identification | 14 |
| 3.2.1 Transactional Data Source | 14 |
| 3.2.2 Digital Data Source | 14 |
| 3.2.3 Operational Data Source | 15 |
| 3.3 Transactional Data Collection | 15 |
| 3.4 Digital Platform Data Collection | 17 |
| 3.4.1 Tracking mechanism for collecting digital data | 17 |
| 3.4.2 Digital attributes | 19 |
| 3.5 Operational Data Collection | 22 |
| 3.6 Data Processing | 22 |
| 3.6.1 Transactional Data Processing | 23 |
| 3.6.2 Aggregation of Transactional Dataset | 27 |
| 3.6.3 Digital Data Processing | 27 |
| 3.6.4 Digital Data Aggregation | 31 |
| 3.7 Operational Data Processing | 33 |
| 3.8 Final Dataset for Predictive Model | 33 |
| 3.9 Summary | 34 |
| **CHAPTER 4: FEATURE SELECTION AND MODELLING** | **35** |
| 4.1 Introduction | 35 |
| 4.2 Feature Selection for Predictive Model | 35 |
| 4.2.1 Outcome of feature selections | 35 |
| 4.2.2 Analysis of feature selections | 37 |
| 4.3 Selection of GBRT Model for Seat Sale Prediction | 37 |
| 4.3.1 Model Selection | 37 |
| 4.3.2 Choosing a Hypothesis Test | 37 |
| 4.3.3 Distribution Estimation | 38 |
| 4.3.4 Hypothesis Testing with Nonparametric Tests | 38 |
| 4.4 Mathematical Model of Gradient Boosted Regression Tree | 39 |
| 4.5 Training Procedure of GBRT and Optimization Using Hyperparameter Tuning | 40 |
| 4.6 Summary | 41 |
| **CHAPTER 5: RESULT AND ANALYSIS** | **42** |
| 5.1 Introduction | 42 |
| 5.2 Analysis of Final Dataset | 42 |
| 5.2.1 Basic Pattern in historical data | 43 |
| 5.2.2 Price change characteristics | 45 |
| 5.3 Result analysis of GBRT Seat Sale prediction model | 46 |
| 5.4 Summary | 51 |
| **CHAPTER 6: CONCLUSION** | **52** |
| 6.1 Conclusion | 52 |
| 6.1.1 Digital Data Integration | 52 |
| 6.1.2 GBRT Model for Seat Sale Prediction | 52 |
| 6.2 Future Work |  |

# LIST OF TABLES

|  |  |  |
| --- | --- | --- |
| [Table 2.1](#_bookmark19) | Summary of previous research on airline demand prediction | 9 |
| [Table 2.2](#_bookmark22) | Parameter used in previous research | 10 |
| [Table 3.1](#_bookmark31) | GIF request types | 19 |
| [Table 3.2](#_bookmark38) | Raw attributes from digital platforms | 20 |
| Table 3.3 | Category of variables from raw digital platform attributes | 22 |
| Table 3.4 | Booking Data | 24 |
| Table 3.5 | Booking Passenger Data | 24 |
| Table 3.6 | Inventory Leg Data | 25 |
| Table 3.7 | Passenger Fee Data | 25 |
| Table 3.8 | Passenger Fee Charge Data | 26 |
| Table 3.9 | Payment Data | 26 |
| Table 3.10 | Sample transactional dataset description | 29 |
| Table 3.11 | Raw Digital Data | 30 |
| Table 3.12 | Sample dataset of the Digital Platform | 32 |
| Table 3.13 | Sample of the final dataset for predictive model | 33 |
| Table 4.1 | Feature Selection result using Boruta | 36 |
| Table 4.2 | Feature Selection result (REF and FFE) | 37 |
| Table 4.3 | Nonparametric Test Result for GBRT-SVM and GBRT-GLM | 40 |
| Table 4.4 | Hyperparameters | 42 |
| Table 5.1 | MAPE pattern for a period of 6 months | 50 |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| [Figure 3.1](#_bookmark28) | Transactional Data Collection processing | 16 |
| Figure 3.2 | Data processing from multiple sources | 23 |
| Figure 4.1 | MAPE Distribution of GBRT Model | 38 |
| Figure 4.2 | MAPE Distribution of GLM Model | 38 |
| Figure 4.3 | MAPE Distribution of SVM Model | 39 |
| Figure 4.4 | GBRT hyperparameters tuning process | 42 |
| Figure 5.1 | The cumulative sum of the total seat sold in different purchase lead days | 45 |
| Figure 5.2 | Average base fare of seat sale in different days (remaining to flight departure) | 46 |
| Figure 5.3 | Price change distribution in last 30 days before a flight departure | 47 |
| Figure 5.4 | Seat sale prediction for 30 days | 49 |
| Figure 5.5 | Seat sale prediction for 60 days | 50 |
| Figure 5.6 | Comparison of accuracy (average MAPE) for 28 routes using proposed GBRT model (traditional only vs addition of digital attributes approach) | 51 |