Predicting e-commerce orders returns and cancellations
using machine learning
Master Thesis Business Analytics
HMN Yousaf

VU supervisor: Prof.dr. Sandjai Bhulai
VU second reader: Dr. Bram Gorissen
PVH supervisor: Nanne Sluis

Vrije Universiteit Amsterdam
Faculty of Sciences
Business Analytics
De Boelelaan 1081a
1081 HV Amsterdam

June 30, 2017
Preface

This thesis is composed as a part of the Business Analytics Master program at the Vrije Universiteit of Amsterdam. It contains the detailed analysis and findings of my internship task at PVH Corp, a company which operates a diversified portfolio of iconic lifestyle apparel brands led by Calvin Klein and Tommy Hilfiger.

The goal of this internship task was to apply the knowledge and latest methods learned from the theoretical work of Business Analytics program on a real business environment. It was an interesting task and I would like to thank Nanne Sluis, my manager at PVH, for introducing me to the company’s work and for his guidance to complete this task successfully. Furthermore, I am much grateful to Alexey Chaplygin, a senior data scientist at PVH, for his directions and explaining me the background knowledge of the task. He provided a lot of insights and together with his useful aid on technical matters, I am able to get good results. I would also like to thank Prof.dr. Sandjai Bhulai, my supervisor from the Vrije Universiteit Amsterdam, for thinking along with my research and giving useful recommendations which helped me to select appropriate models for this research. Finally, I would like to thank Dr. Bram Gorissen for being my second reader.
Abstract

E-commerce business is booming everyday as internet has made the world a smaller place and facilitated long distance communication by making the process cheaper, faster, and easier. Fashion business houses have maintained websites to promote and sell their products. It is very common in online fashion business to expect the order returns as customers do not get chance to try them on before receiving those orders. The order returns can be irritating for online sellers as it costs a lot of time and money to process them. This research is completed to help PVH, a company that owns and operates the iconic lifestyle apparel brands led by Calvin Klein (CK) and Tommy Hilfiger (TH), in predicting e-commerce orders returns and cancellations. The machine learning models and techniques are applied to predict the TH and CK orders returns and cancellations with more than 95% accuracy. This study helped PVH to report rolling net sales accurately and plan marketing budget with high accuracy to push more campaign accordingly.
Contents

Preface .............................................................................................................. 1
Abstract ............................................................................................................. 2
1 Introduction ............................................................................................. 4
  1.1 Purpose of this research ................................................................. 4
  1.2 Paper Overview ............................................................................... 5
2 Background & Literature Review ............................................................ 6
3 Data Extraction, Description and Preprocessing ....................................... 8
  3.1 Date Extraction ................................................................................ 8
  3.2 Data Description .............................................................................. 9
  3.3 Data Exploration and Features Engineering ...................................... 10
    3.3.1 TH ........................................................................................... 10
    3.3.2 CK ............................................................................................ 11
    3.3.3 Features Engineering .................................................................. 13
4 Modeling Methods ..................................................................................... 19
  4.1 Models ............................................................................................. 19
    4.1.1 Random Forest ...................................................................... 19
    4.1.2 Glmnet .................................................................................... 19
  4.2 Evaluation ......................................................................................... 21
  4.3 R Packages and Procedure ............................................................. 23
5 Results ..................................................................................................... 25
  5.1 TH ................................................................................................... 25
  5.2 CK ................................................................................................. 30
6 Conclusion & Discussion .......................................................................... 36
References


