Cab Booking Cancellations

By: Nivetha Sivaprakasam

Overview

- Introduction
- Dataset
- Wrangling
- Logistic Modeling
- Weight of Evidence(WOE)/Information Value(IV)

Overview cont.

- Decision Tree
- Information Gain
- Results
- Recommendation
- Future

Introduction

- Customers rely on cabs as a method of quick transportation
- Problems arise due to lack of cabs
- Customers are unhappy and panicked to search for a new cab
- Goal: Explore possibilities to decrease cab cancellation chances

Dataset

- Kaggle’s in class competition “Predicting cab booking cancellation”
- Package_id: (1 = 4hr & 40 kms, 2 = 8hr & 80kms, 3 = 6hr & 60kms, 4= 10hr & 100kms, 5 = 5hr & 50 km, 6 = 3hrs & 30 km, 7 = 12hrs & 120 kms)
- Travel_type_id: (1 = long distance, 2 = pt to pt, 3 = hourly rental)
- Rest of variables are binary (1 = true, 0 = false)

Wrangling

- Adding package_id value of (0 = NA values)
- Separating from_date & to_date to month, day, date (from_month, from_day, from_time, to_month, to_day, to_time)
- New variable diffs - difference in days rounded from booking_created and from_date

Logistic Regression

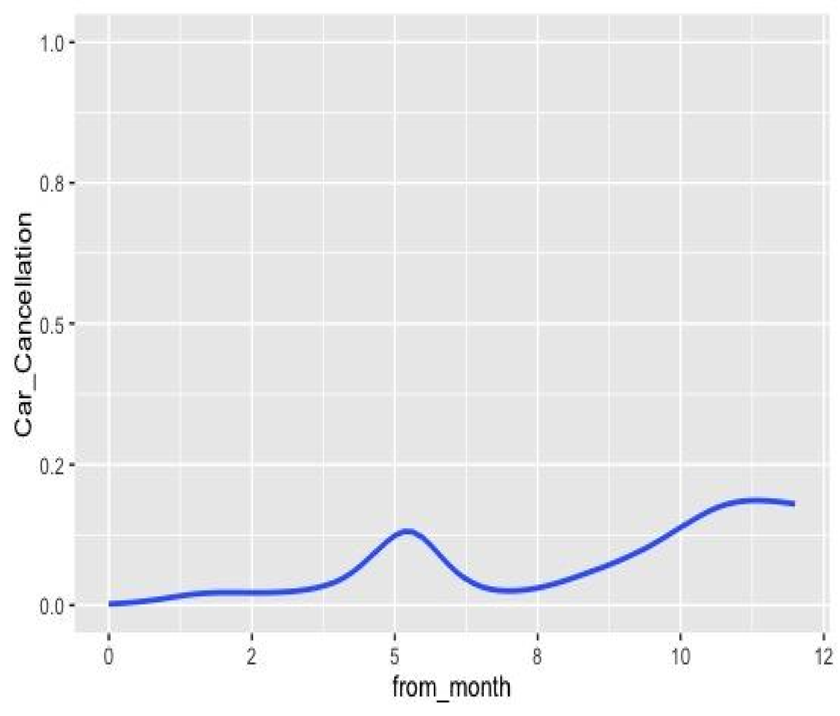
- Is a statistical method to analyze data in which there is 1 or more independent variable determining an outcome.
- Need to find important variables which are done with Weight of Evidence (WOE) and IV (Information Value)
- IV is used to see the predictive power of the variables.

Logistic Regression cont.

- IV value of < 0.02 USELESS
  - 0.02 to 0.1 WEAK
  - 0.1 to 0.3 MEDIUM
  - 0.3 to 0.5 STRONG
  - 0.5 > SUSPICIOUS
- Sample Run of iv.mult (function used to determine IV and WOE

Information Value 0.31

Curve graphs



Logistic Curve for from_month

Increased chance of cancellation around May & November

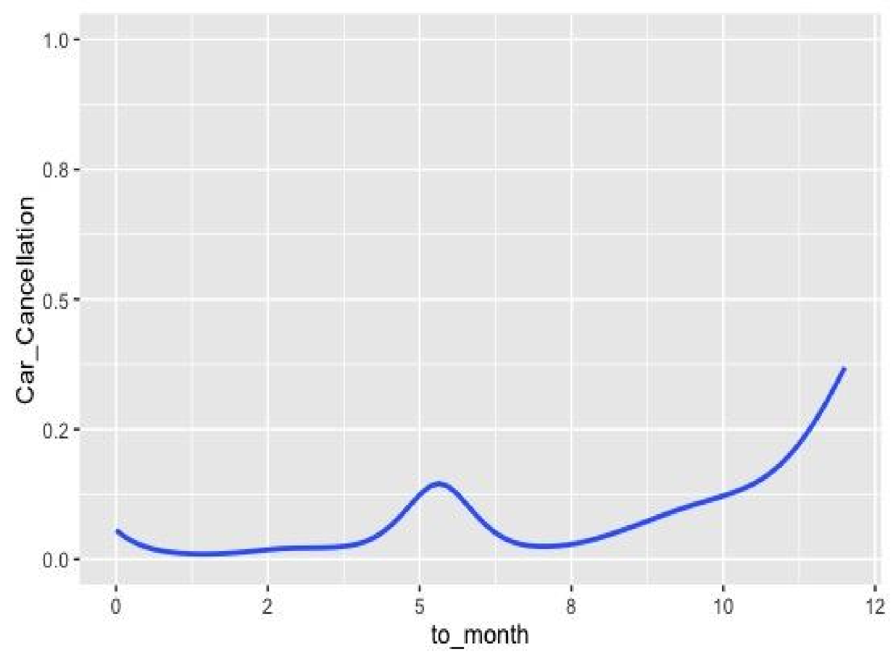
Decreased chance around February & July

Curve Graphs cont

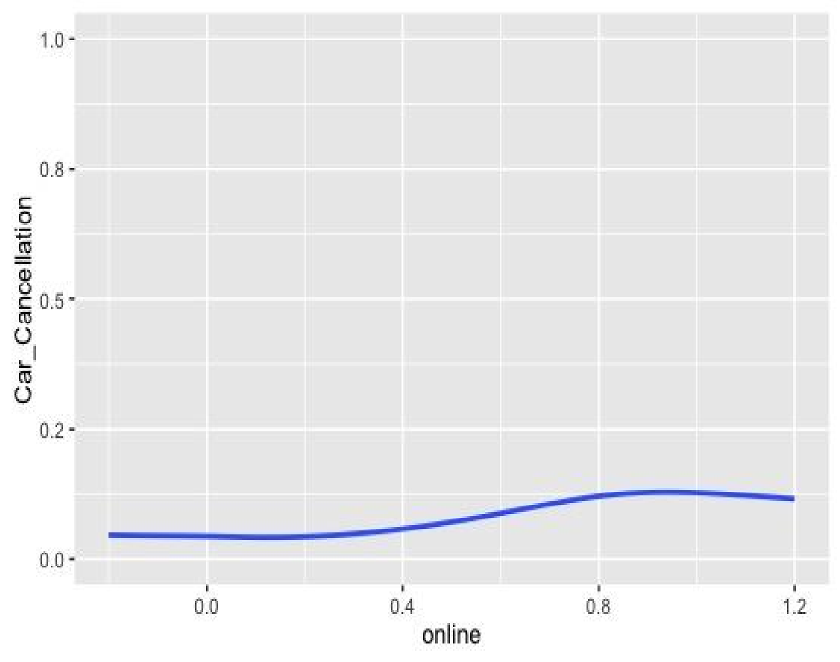
Logistic Curve for to_month

Higher cab cancellation occurs June, October & November

Lower cab cancellation occurs January and July



Curve Graphs cont



Logistic curve for online

Higher cab cancellation chances occur when a booking is done online

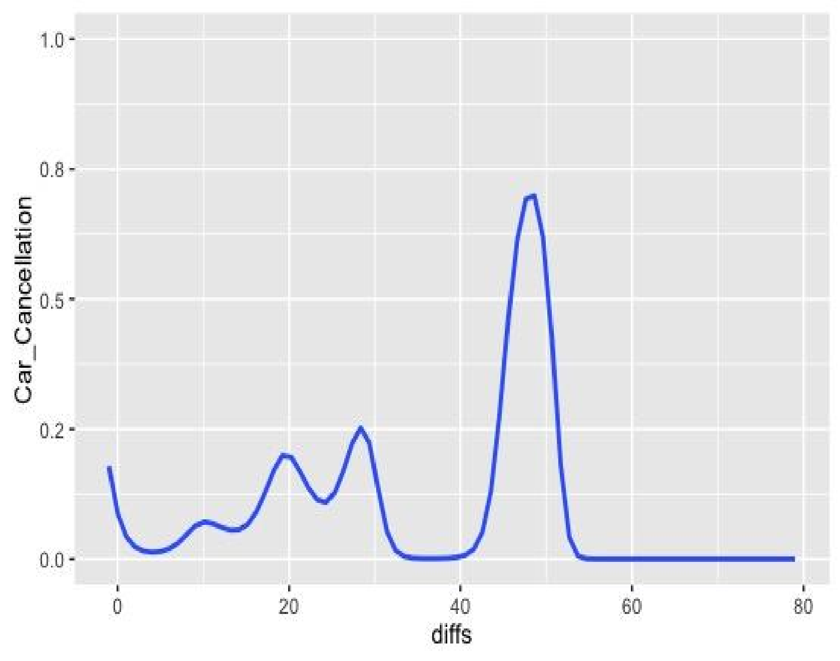
Lower cab cancellation occur when is booking is NOT done online

Curve Graphs cont

Logistic Curve for diffs (Difference in days from booking_created and from_date)

Higher chances of cab cancellation occur during 50 and 30 days

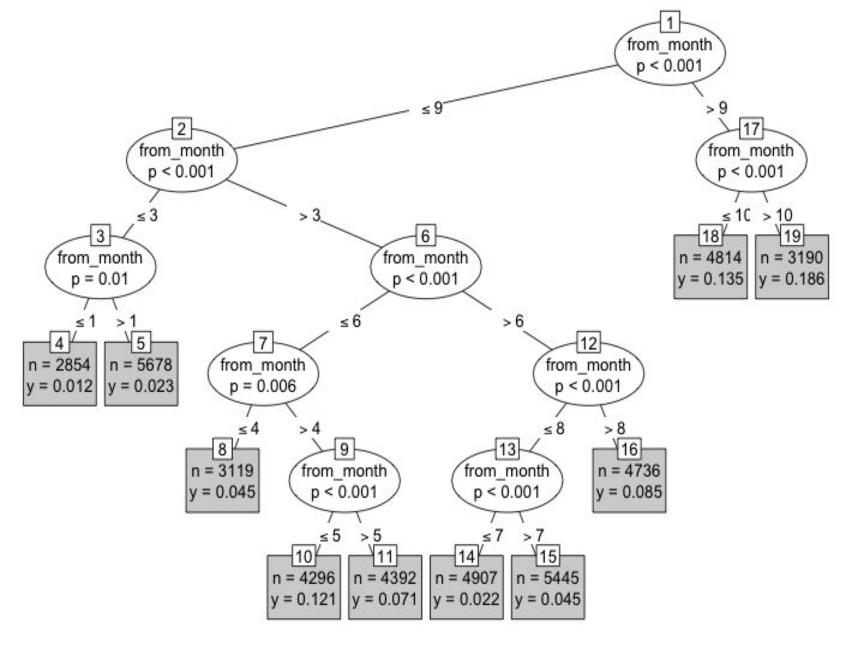
Lower chance of cancellation occur during 5 days and 60 to 80 days



Decision Tree & Information Gain

- Tree like graph that uses decision model to visualize possible consequence and number of nodes
- Information gain is used to see how much information is gained by splitting by the variable
- Used the party package and ctree function call

Decision Tree cont



Decision Tree for the variable from_month

Popular month includes January

Highest cancellation chance occurs after October and May

Lowest cancellation chance occurs during January and July

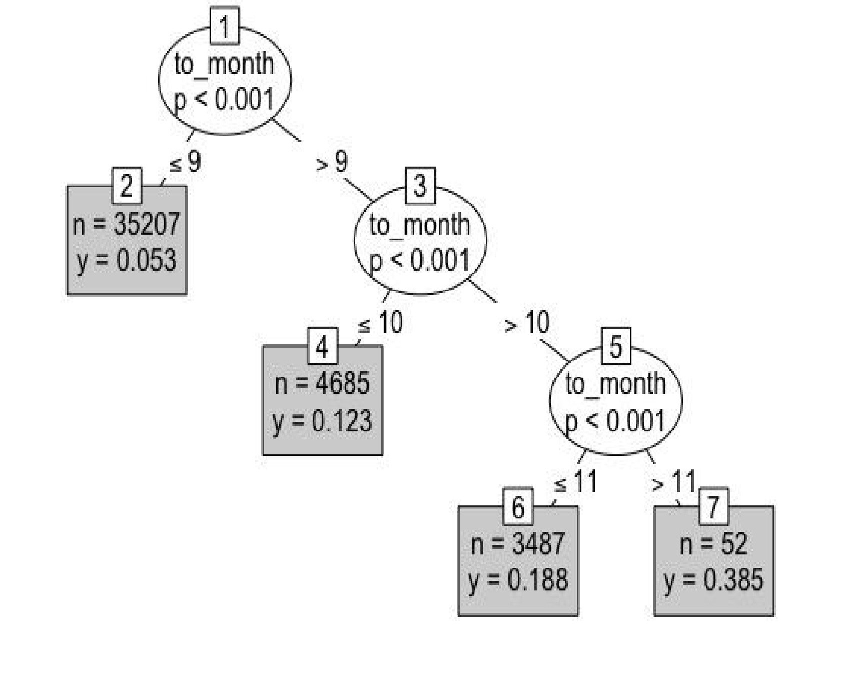
Decision Tree cont

Decision Tree for to_month

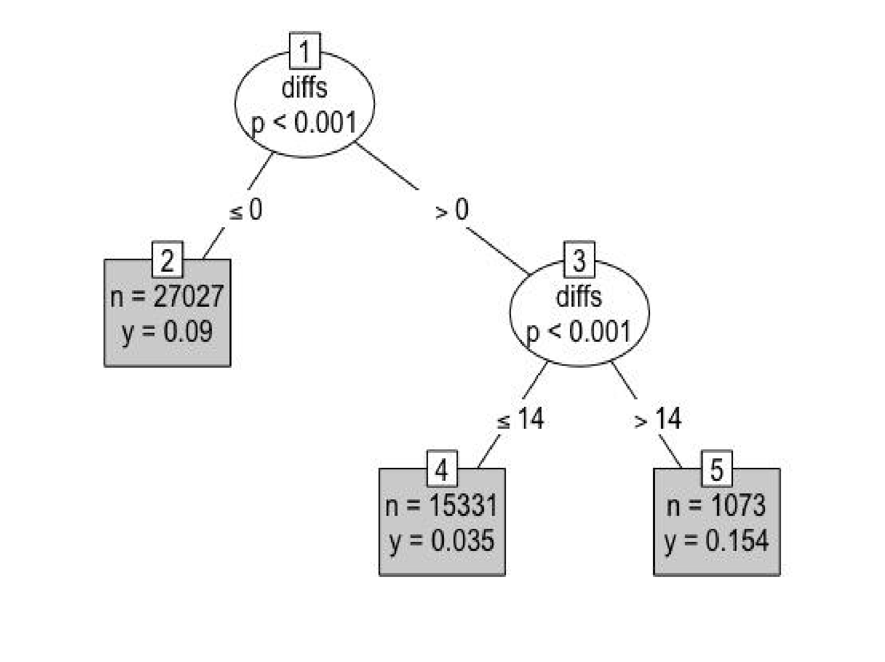
Popular months are before September

Higher cab cancellations occur during December

Least cab cancellations occur before September



Decision Tree cont



Decision Tree for diffs

Common Values are for 0 days

Higher chances for cancellation occur for difference greater than 14

Least chance for cancellation occur for difference less than 14 but greater than 0

Result

- Mobile and online bookings have greater chance of cancellation
- A package id of 7 (12hrs & 120 km) have great chance of cancellation compared to 3(6hrs & 60 km)
- Long distance is less cancellation compared to point to point travel

Future/Recommendations

- Companies
  - Prioritize early morning and late night customers for safety concerns
  - Should give early notice for cancellation
  - Stock up on more cabs and give other options to customers
- Customers
  - Prepare for a back up plan
  - Avoid online & mobile booking or double check with company via phone
- Research
  - Consider different countries
  - Include distance travelled as a variable

THE END