Forecasting bus ridership with trip planner usage data

a machine learning application

Jop van Roosmalen

Acknowledgement:
Dr. Chintan Amrit (UTwente)
Dr. Engin Topan (UTwente)
Dr. Niels van Oort (Smart Public Transport Lab)
9292 Trip planner

Waar wil je heen?

Locaties:
- Gouda
- Hallenweg 17, Enschede

Datum: 28-03-2019
Tijd: 15:45

Type reis:
- Vertrek
- Aankomst

Plan mijn reis ➔

Vertrek: Spoor 3
Station Gouda

Aankomst: Spoor 3
Station Hengelo

Lopen
3 minuten

Aankomst
Bushalte Centraal Station, Hengelo
Vertrek, Perron A1

Syntus Stadsbus 9
Richting Enschede C6
Introduction

Objective

• Construct a forecasting model
• Determine the accuracy of the models
• Investigate predictive power of trip planner usage data
• Determine valuable features
Methodology

Models

• $Passenger_{stop} = Passenger_{stop-1} + Boarding_{stop} - Alighting_{stop} = \sum_{i=0}^{s} B_i - \sum_{i=0}^{s} A_i$

Machine learning

• Multiple linear regression
• Decision tree - decision tree regressor
• Random forests
• Support vector regression with radial basis kernel
• Artificial Neural Networks - Multi-layer Perceptron regressor

Comparison with simple rules

1. Predicted number equals number last week
2. Predicted number equals historical average
Methodology
Undersampling using stratified K-fold

Stratified k-fold

Under sampling of training data

Classes

k=1

k=2

k=.. (ellipsis)

k=10

Nonzero | Zero | Neglected | Test | Train

[Diagrams showing the process of undersampling with stratified K-fold]
Methodology

Performance metrics

• $RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$

• $R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$

• % of passenger count predictions correct
• % of maximum passenger count predictions correct

• Python, Scikit-learn
Case study

Scope

• Data from Groningen and Drenthe
• 4,972 km² Land area
• ± 1.1 mil Habitants
• ± 0.2 mil Habitants Groningen City
• January to March 2017
• Time period contains two smaller holidays
Data Structure

Trip planner
11,694,849

Smart card
6,814,907
4,946 stops

AVL data
11,447,562

All on vehicle level
Data

Merging trip planner with bus data

• 6 – dimensional problem
• Almost no exact matches!

Metric:

*Difference boarding times + difference alighting times*
Data

Exploratory data analysis
Data
Exploratory data analysis
Data

Data selection

Forecasting demand for trips of line configuration g554-1-0 on workdays around 8 AM

1. 20 lines on workdays around 8 AM
   (56 line configurations, 4173 trips and 138,694 records)

2. 20 lines configurations for the total workday
   (83 line configuration, 51,471 trips and 1,523,115 records)

3. line configuration g554-1-0 for the total workday
   (1 line configuration, 2275 trips and 97,825 records)

4. line configuration g554-1-0 on workdays around 8 AM
   (1 line configuration, 239 trips and 10,277 records)
Data
Line configuration g554-1-0

- From Roden via P+R and Groningen central Station to Hospital
- 43 stops
- 631 m average stop spacing
- 26 km total route (partly own lane)
- 61 minutes from begin to end
- 6-2 busses an hour
Results

RMSE

MLR

DT

RF

NN

SVR

Last week

Historical avg
Results

RMSE Passengers
Results
Passenger prediction example

• g554-1-0
• Trip 1018
• February 15, 2017
• Wednesday
• 07:22 – 08:26
Results
Percentage correct maximum passenger count predictions

1. Last week
2. Historical average
Random Forests
Discussion

Limitations

• One trip planner, no session id
• Only smart card
Conclusion

Research question

*Can one forecast short-term ridership of buses using data containing the consulted travel advices from a widely used trip planner for public transport and what accuracy can one achieve in different scenarios?*
**Practice**

- Adapt data structure for data analysis
  - Include bus trip number, line number, operation date and stop
  - Include session ID
  - Trip level
  - Use same set of stops
- Models

**Research**

- Forecasting structure

  - Features:
    - Which
    - Form
    - Scaling
    - Amount
  - Training data:
    - Size
    - Quality
  - Performance metric:
    - Average
    - Upper bound
  - Models:
    - Type
    - Complexity
    - Running time
    - Tuning
      (bias/flexible)

Forecasting performance
Thanks for your attention

jop.j@hotmail.com
linkedin.com/in/jop-van-roosmalen/
Slides nielsvanoort.weblog.tudelft.nl
Thesis essay.utwente.nl/77590/