AUTOMATICALLY FILLING OUT DRIVING JOURNAL USING MACHINE LEARNING ALGORITHMS

Getter Põru, Jaagup Russak
University of Tartu, Estonia

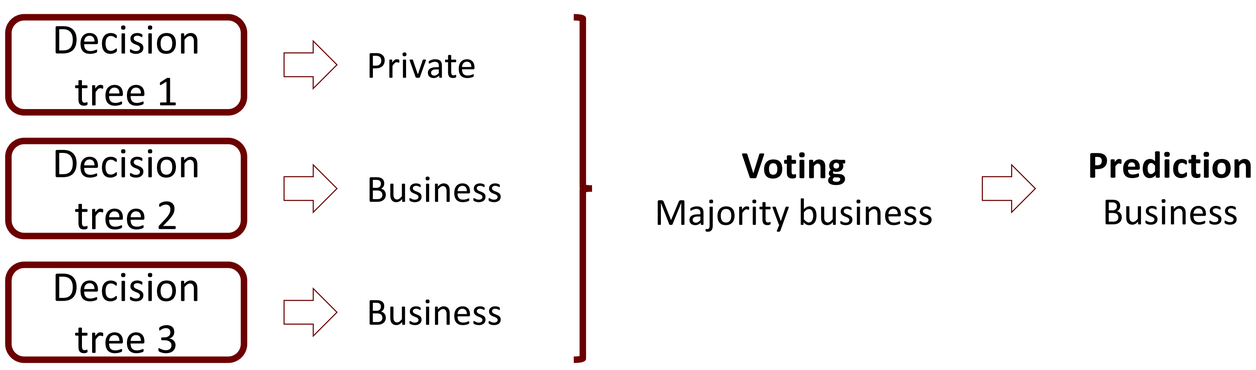
https://bitbucket.org/jrussak/datascience

INTRODUCTION

It is very common that organizations allow their employees to make business trips with their personal cars and use their business cars for personal matters. Then after filling out a report each month they get some refund on the business trips. However filling out these reports is a long and boring process and it takes a lot of unnecessary time. The aim of this work is to use machine learning algorithms to fill out these reports for them automatically.

DATA

The dataset contains around 6 million trips from 2853 Denmark cars. Each trip has a manually added purpose (business or private) and additional features like GPS coordinates, date, time, distance and duration. About 68% of the trips are business trips and 32% are private trips.



METHODS

A **random forest** is a data construct applied to machine learning that develops large numbers of random decision trees analyzing sets of variables. As there are some **categorical variables** in this dataset we have used **H2O** software for model fitting which does not require the data to be one-hot encoded and maintains the relationship between different levels. **PCA (Principal Component Analysis)** is a statistical method that converts a set of observations of correlated variables into a set of values of linearly independent and therefore uncorrelated variables (principal components **PC**).

Fig. 1. Ending points of the first 100 000 trips.

RESULTS AND DISCUSSION

The highest accuracy achieved on the test data was **0.94**. Among the best models experimented with were models which included principal components (distance, duration, coordinates, time) as features. We also tried creating a separate model for each car but that did not lead to a higher accuracy.

As the highest accuracy was achieved with using the highest computational power needing algorithm parameters we tried, it is reasonable to believe that a better computer would increase the accuracy of the model. Also some segmentation like city based or organization's type based models could be tried out in the future work.

Acknowledgements
The data used in this project was provided by Fleet Complete.