Unemployment Length Prediction

Marcelo Bohrer

12/19/2016

Introduction

The idea of this project is to develop an algorithm to predict the length of unemployment in the US economy based on some economic indicators.
Predicting the length of unemployment helps:

- Optimize government spending on social security.
- Predict debt default rates.
- Understand the Economy as a whole

The Data

Part of data is available in the economics data set from ggplot2. Additional variables were obtained from governmetal and international organizations websites.
The variables are [variable codes in brackets]:

- Median weeks of unemployment [uempmed];
- Unemployment rate [urate];
- Consumer price index [cpi];
- Producer price index [ppi];
- Fed interest rate [InterestRate];
- Oil price (month avg.) [OilPrice];
- Month [month];
- Stock Index (Dow Jones) [DowJones];

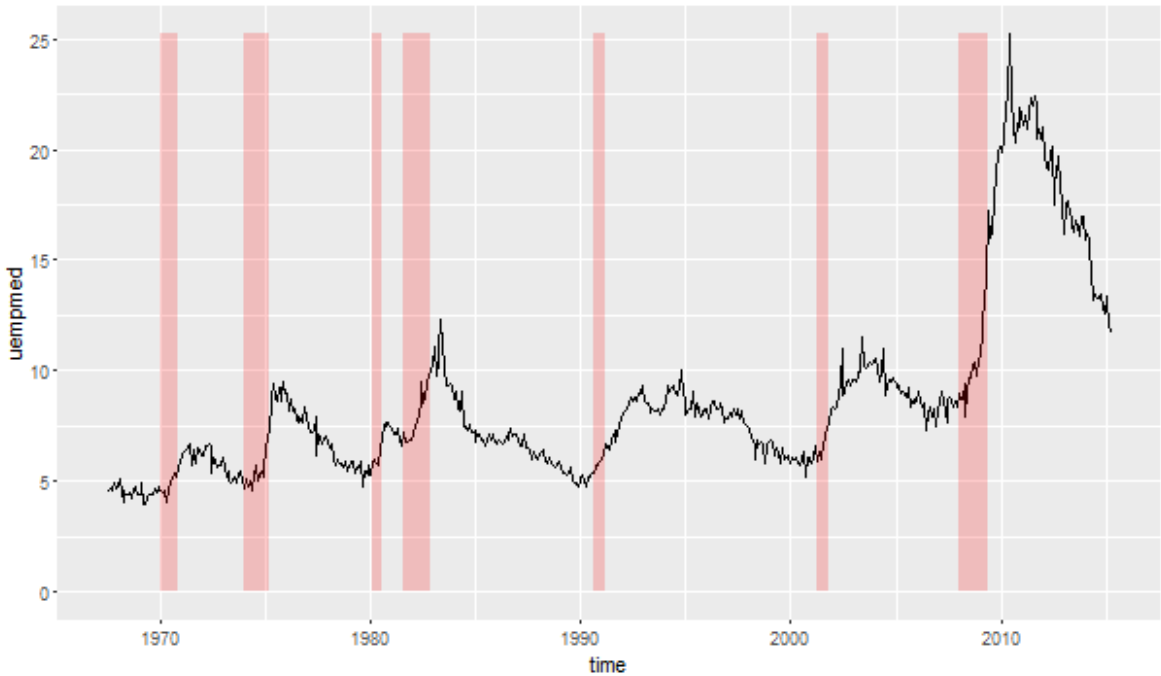
The Data

- Personal savings rate [psavert];
- Personal consumption spending [pce];
- Manufacturing capacity utlization [cap];
- Manufacturing Confidence report [mconf];
- Dummy variable for recession period [USREC].

The data covers the 1967-2015 span, with monthly data collection.

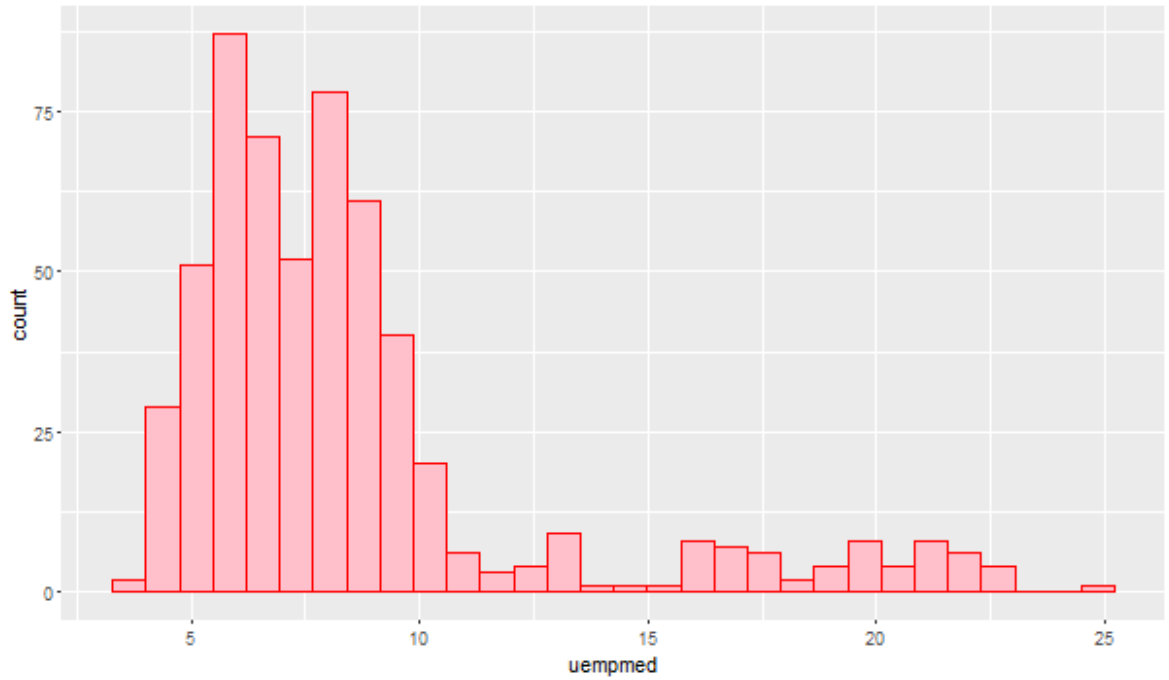
Exploratory Data Analysis

The next plot shows the evolution of the median weeks of unemployment with the recession periods marked.



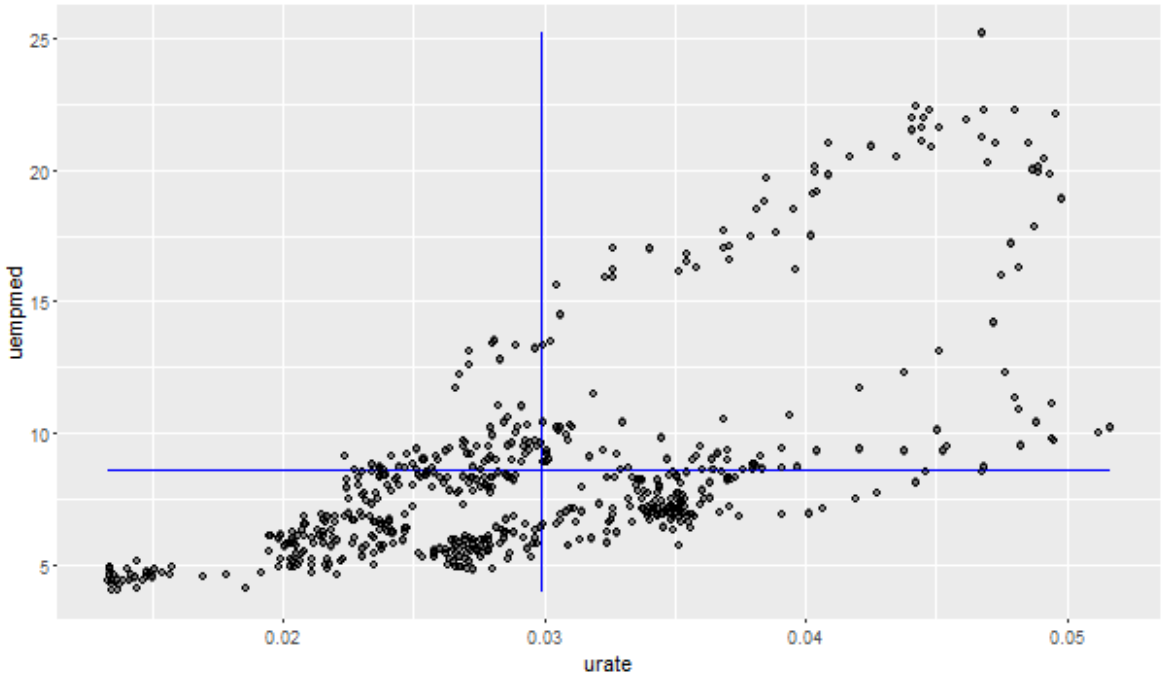
Exploratory Data Analysis

We can analyze the distribution of the median weeks of unemployment.



Exploratory Data Analysis

We can also see how the rate and the length of unemployment are related.



Model Building

We fit 5 different models to the data, namely:

- Linear Regression
- Tree
- Random Forest
- Cubist
- Support Vector Machines

Different models are compared based on their RMSE.

Results

Model	RMSE
lm	1.479
tree	1.412
rf	1.169
cubist	1.128
svm	1.657

The variable importance for the cubist model is as follows:

Variable	Overall
pce	100.00
urate	98.780
cpi	79.878
ppi	51.829
OilPrice	47.561
cap	42.073
mconf	40.244
DowJones	32.317
psavert	20.732
InterestRate	9.756
month	0.000
USREC	0.000

Conclusion

The best fit was achieved by the Cubist model.
The model is off by 1.3 weeks in prediction. The range of our independent variable is roughly 21 weeks, hence a 1.3 error is not too large.