24.05.2021

QPE studies

Responsible Machine Learning

Krzysztof Osesik

**Project**

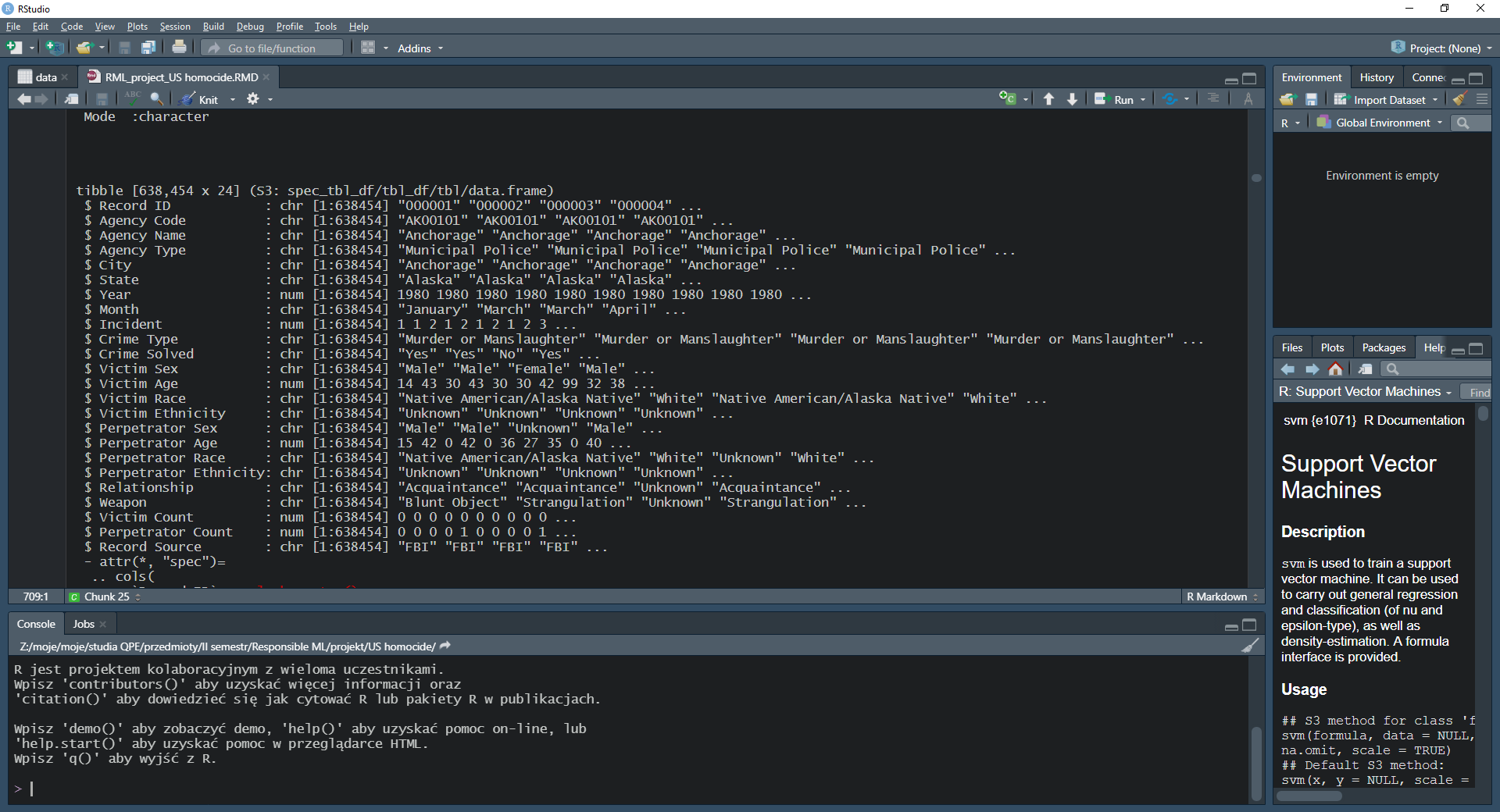
**Introduction**

**Data**

The Murder Accountability Project is the most complete database of homicides in the United States currently available. This dataset includes murders from the FBI's Supplementary Homicide Report from 1976 to the present.

Dataset comes from Kaggle (https://www.kaggle.com/murderaccountability/homicide-reports).

The dataset consists of nearly 640,000 crime cases. They are described by 24 variables, for example City, State, Year, Crime Type, Victim's Age, Perpetrator's Age and Weapon Used.



**Stakeholder**

A stakeholder could be associated with a public authority aiming at ensuring that – in a case of a murder – identical effort is put into investigation with no regard for the victim’s race, sex or age. As a result similar crimes should be solved (alternatively not solved) independent of the victim’s race, sex and age.

The stakeholder is thus interested in evaluating the model’s general performance. Their goal is to analyze whether there is some bias in police investigations towards certain social groups, which effectively results in a lower rate of solved crimes for those groups.

**Target variable**

* Crime Solved

**Explanatory variables**

* Agency Type (e.g. County Police, State Police, etc.)
* Victim’s Sex (Female, Male)
* Victim’s Race (e.g. Black, White, Native, etc.)
* Victim’s Age ( continuous 18-100 years old)
* Weapon (e.g. Knife, Gun, Poison, etc.)

**Protected variables**

* Victim’s Sex
* Victim’s Race
* Victim’s Age

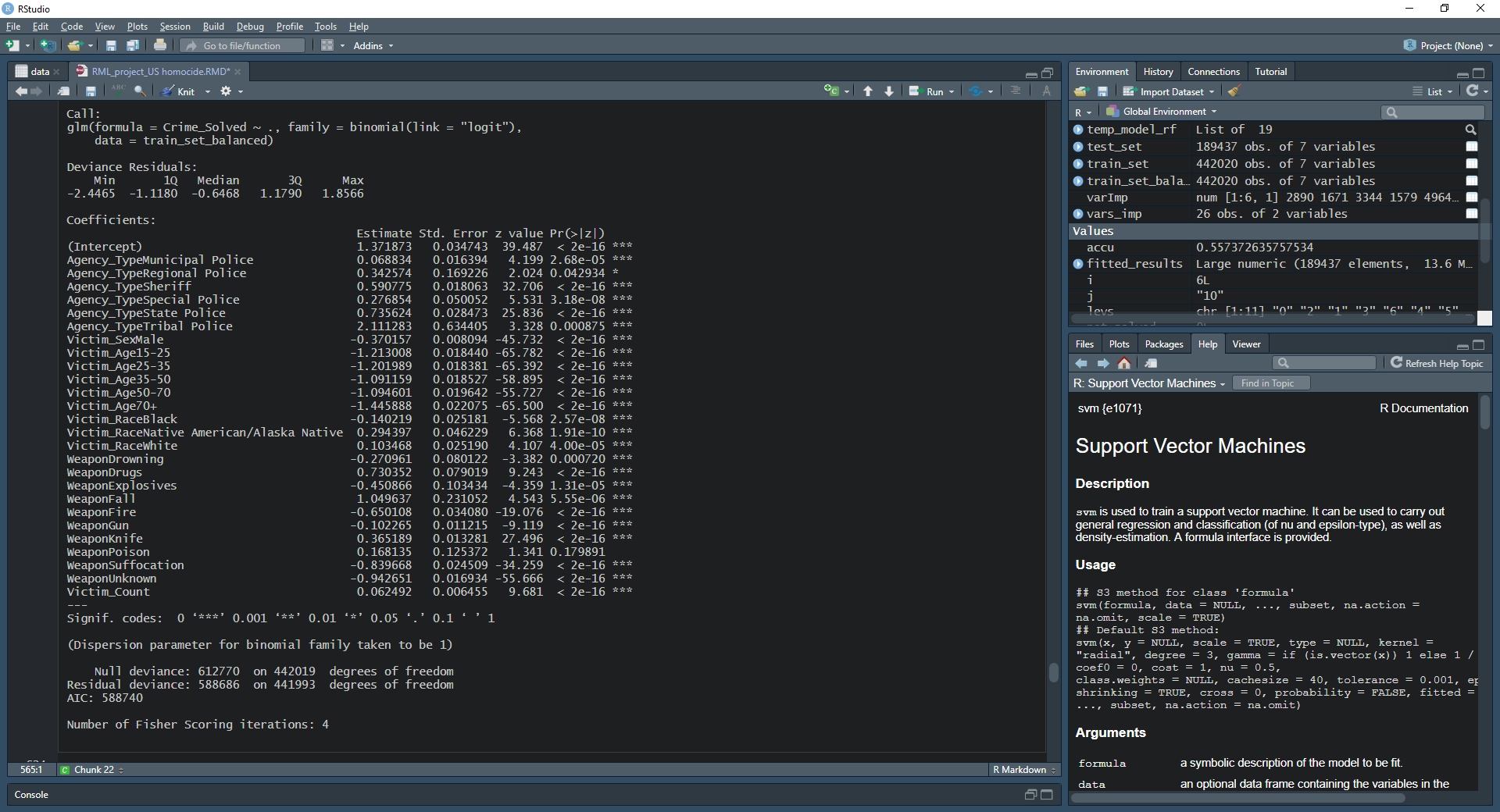
**ML algorithms and variables**

I plan to use two machine learning algorithms:

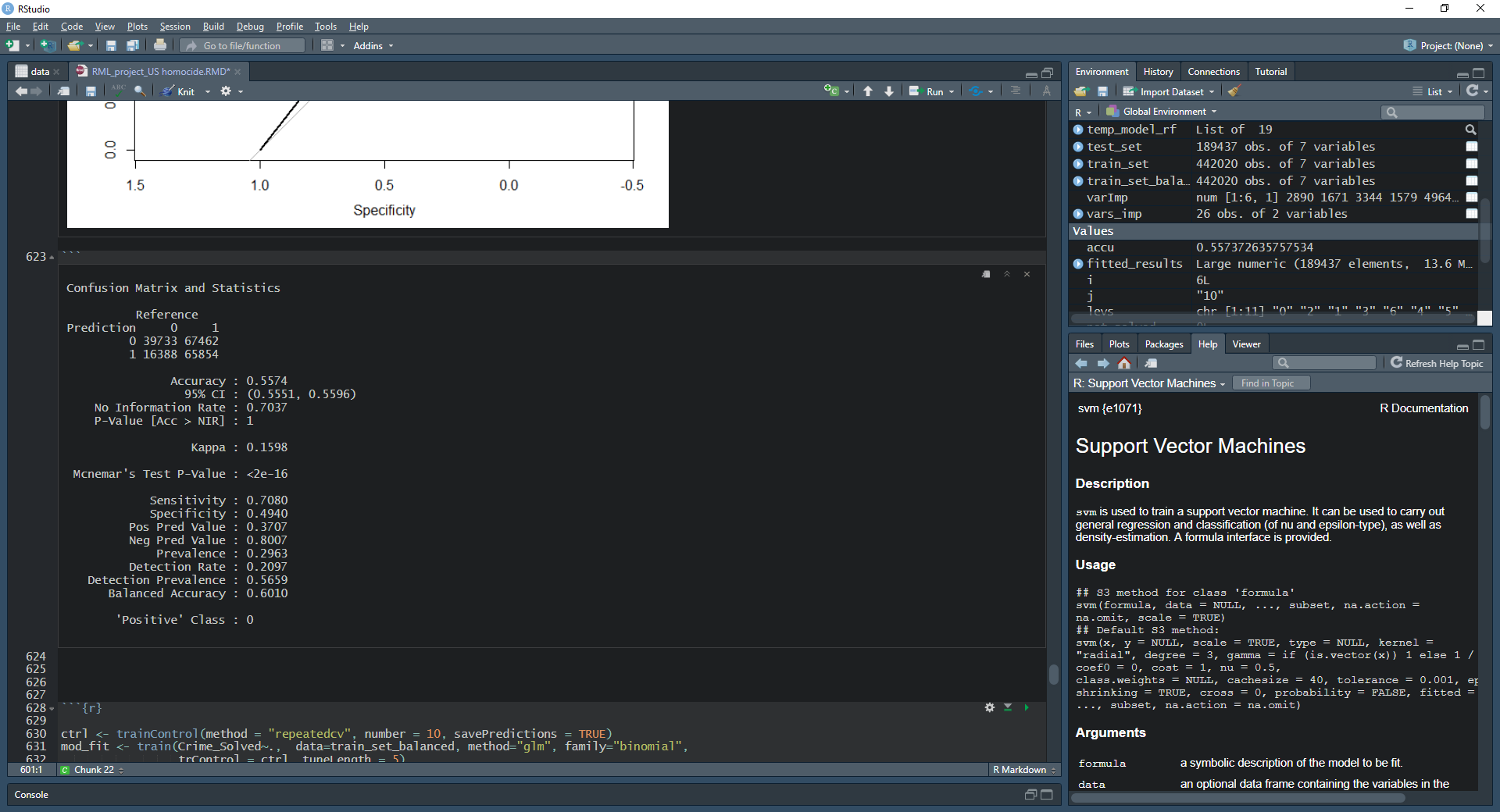
1. LogisticRegression
2. Random Forest

Below I present some initial results.

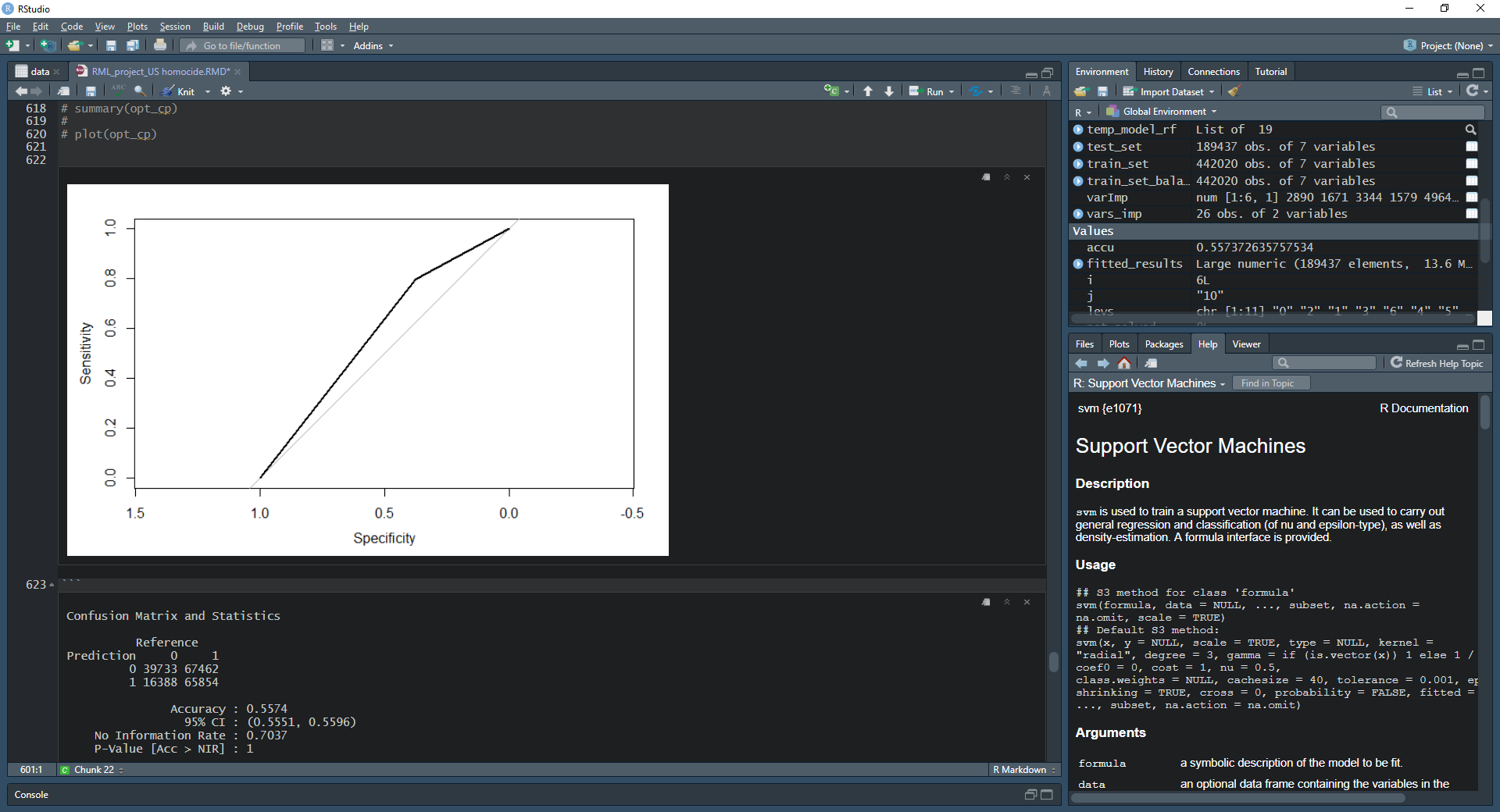
1. Logistic regression



I received the following confusion matrix ( in a confusion matrix -1 denotes that crime was solved and 0 denotes that crime was not solved).



… and ROC plot.

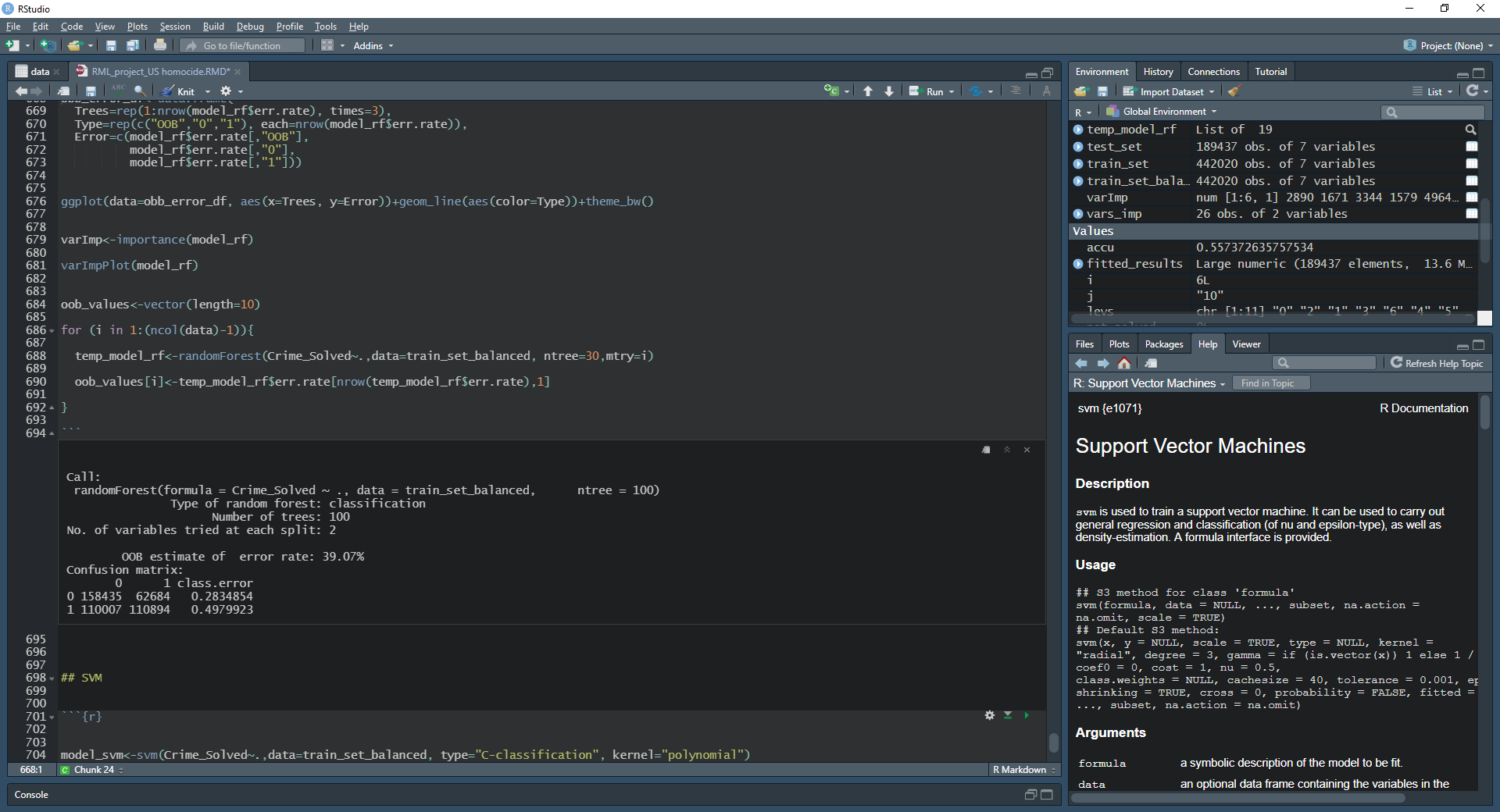


Accuracy: 0.5857

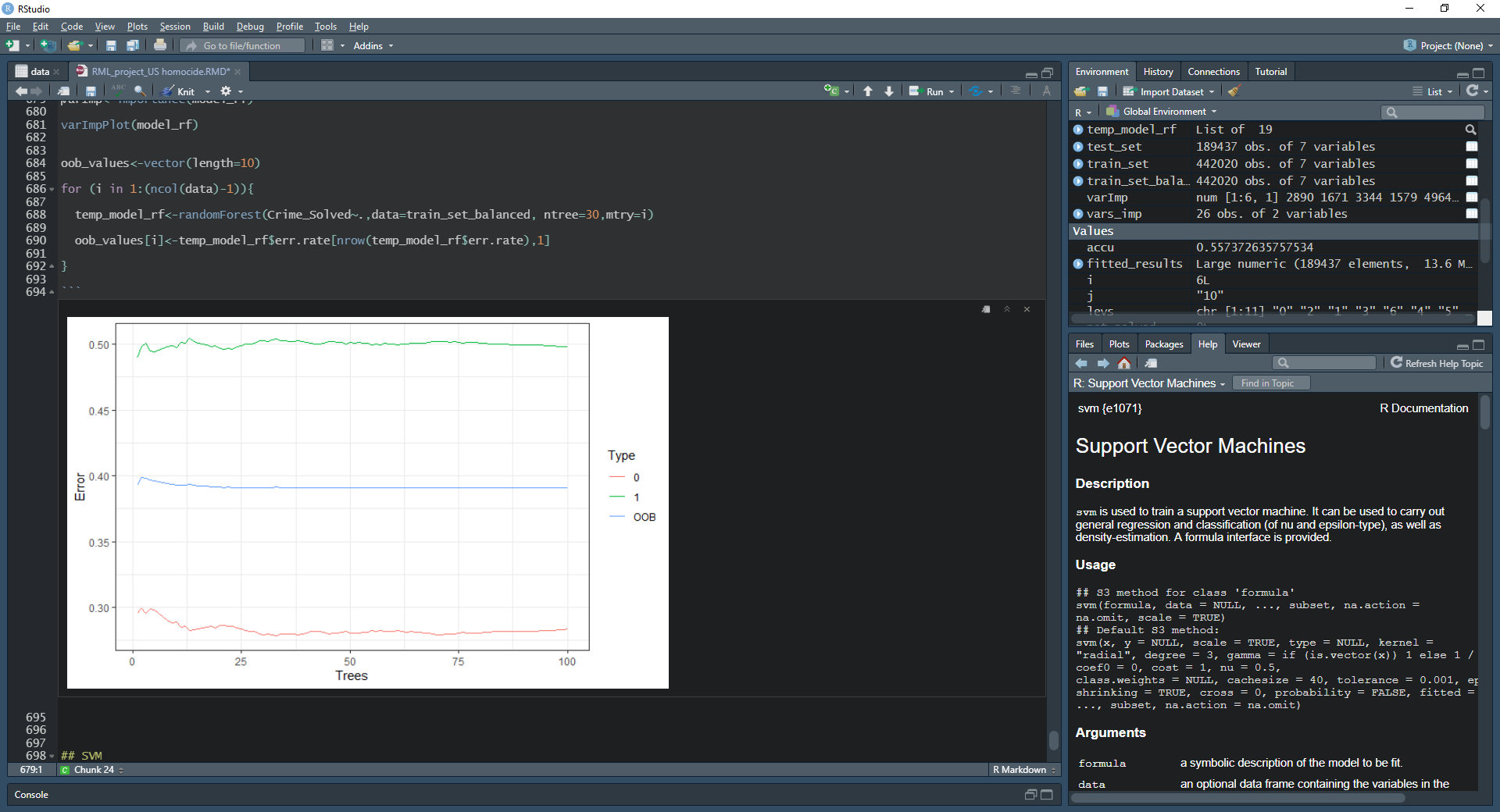
Clearly, the results are poor. I tried to improve them by running 10-fold cross validation on the training set, but the results were the same as with the simple logistic regression.

1. **Random Forest**

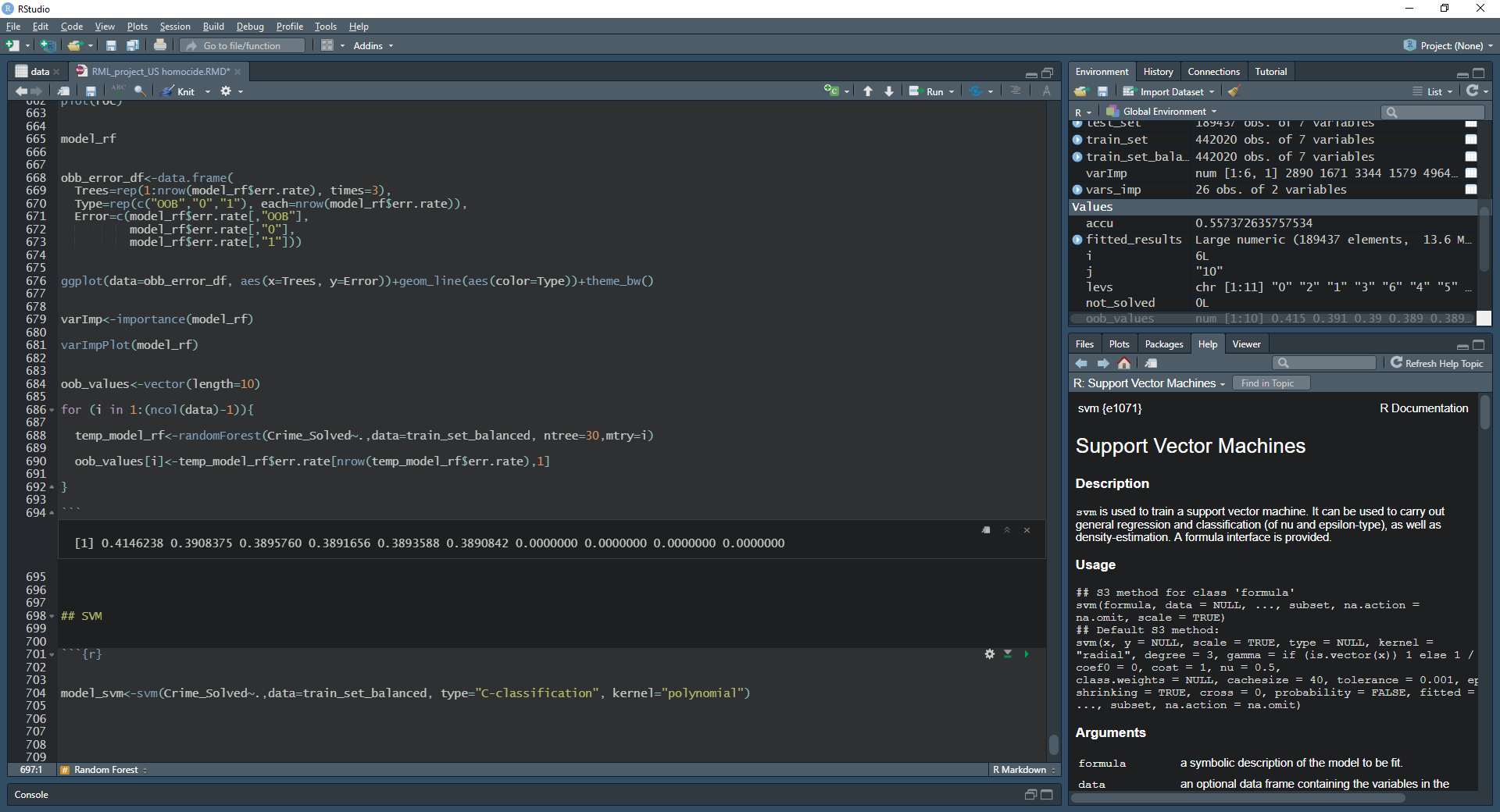
I have run the following analysis**.**



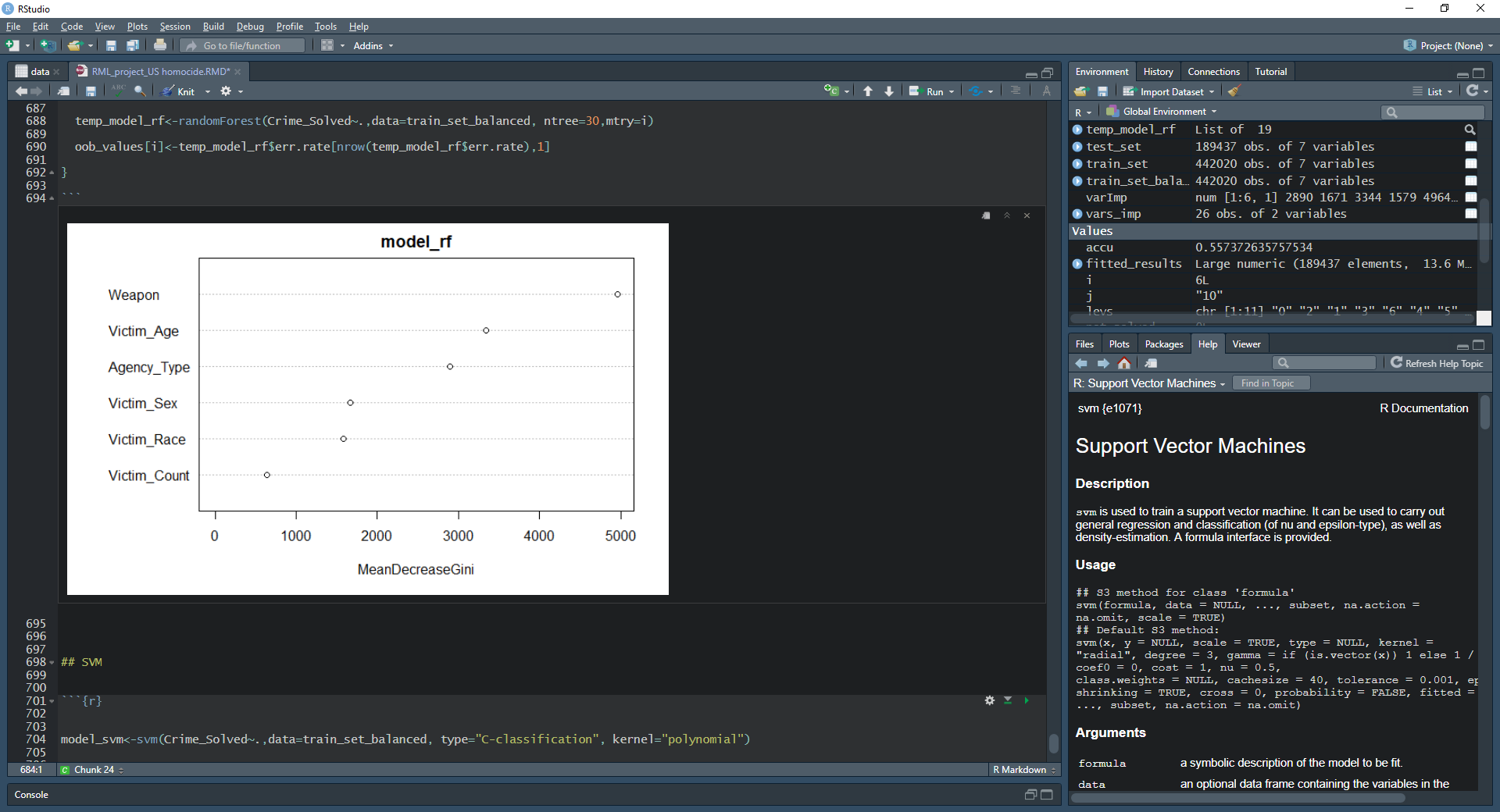
I used 100 trees in the random forest formula, as error rates are stable then.



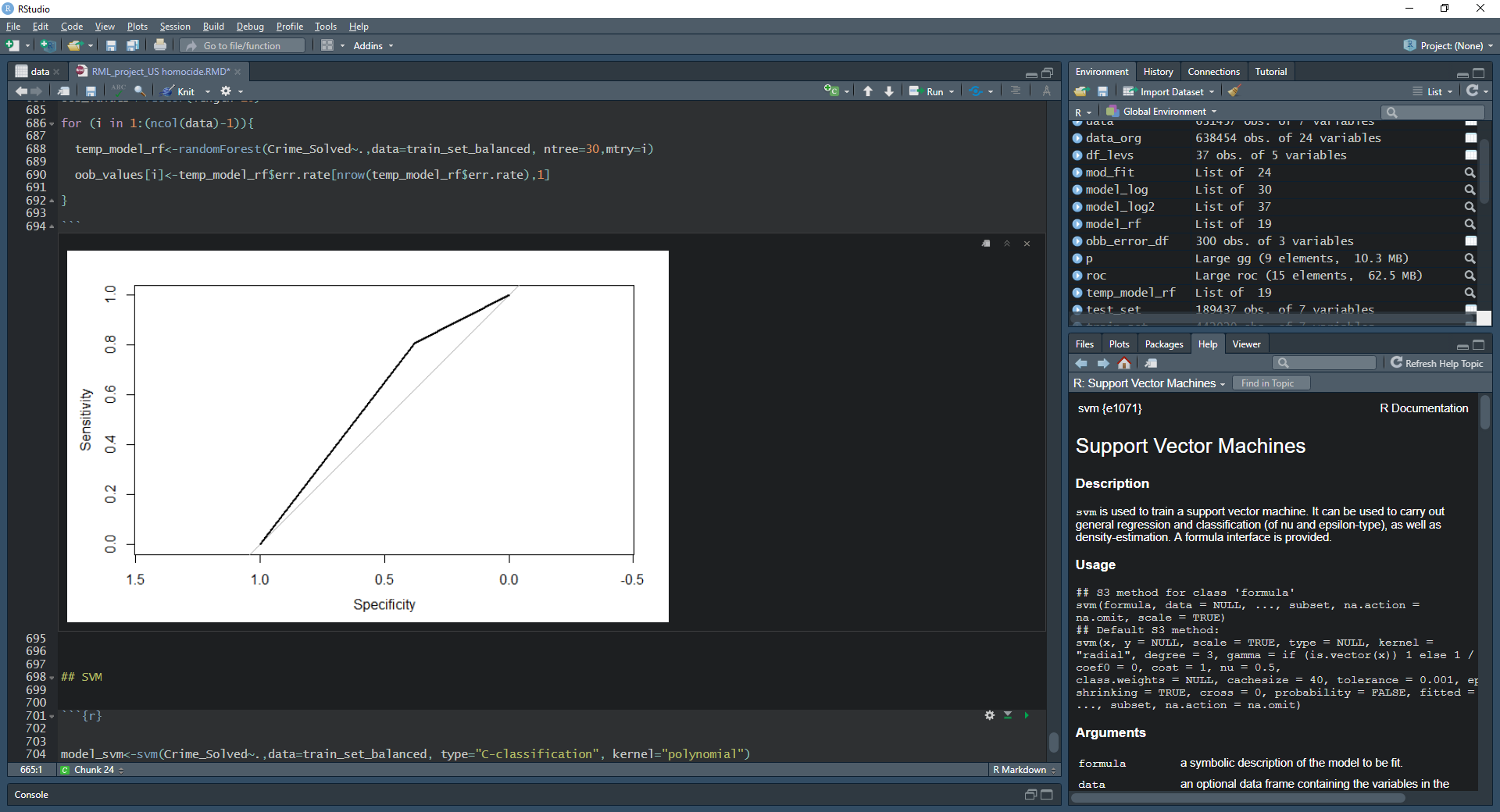
Out-of-bag error depending on the number of variables evaluated at each step of random forest (from 1 variable to max 6 variables). Error rate does not improve almost at all after 2 variables.



Variable importance plot below.



ROC plot is the same as for logistic regression with only slight improvement for accuracy.



Accuracy : 0.5923

**Methods to analyze explainability and fairness**

In order to analyze explainability and fairness of the model I intent to use the following methods:

**Explainability:**

* Permutation Variable Importance
* Partial Dependence Plots
* Shapley Values
* LIME

**Fairness:**

* Unawareness (of the model towards protected variables)
* Demographic parity
* Equalized odds