24.05.2021

QPE studies

Responsible Machine Learning

Krzysztof Osesik

**Project**

**Introduction**

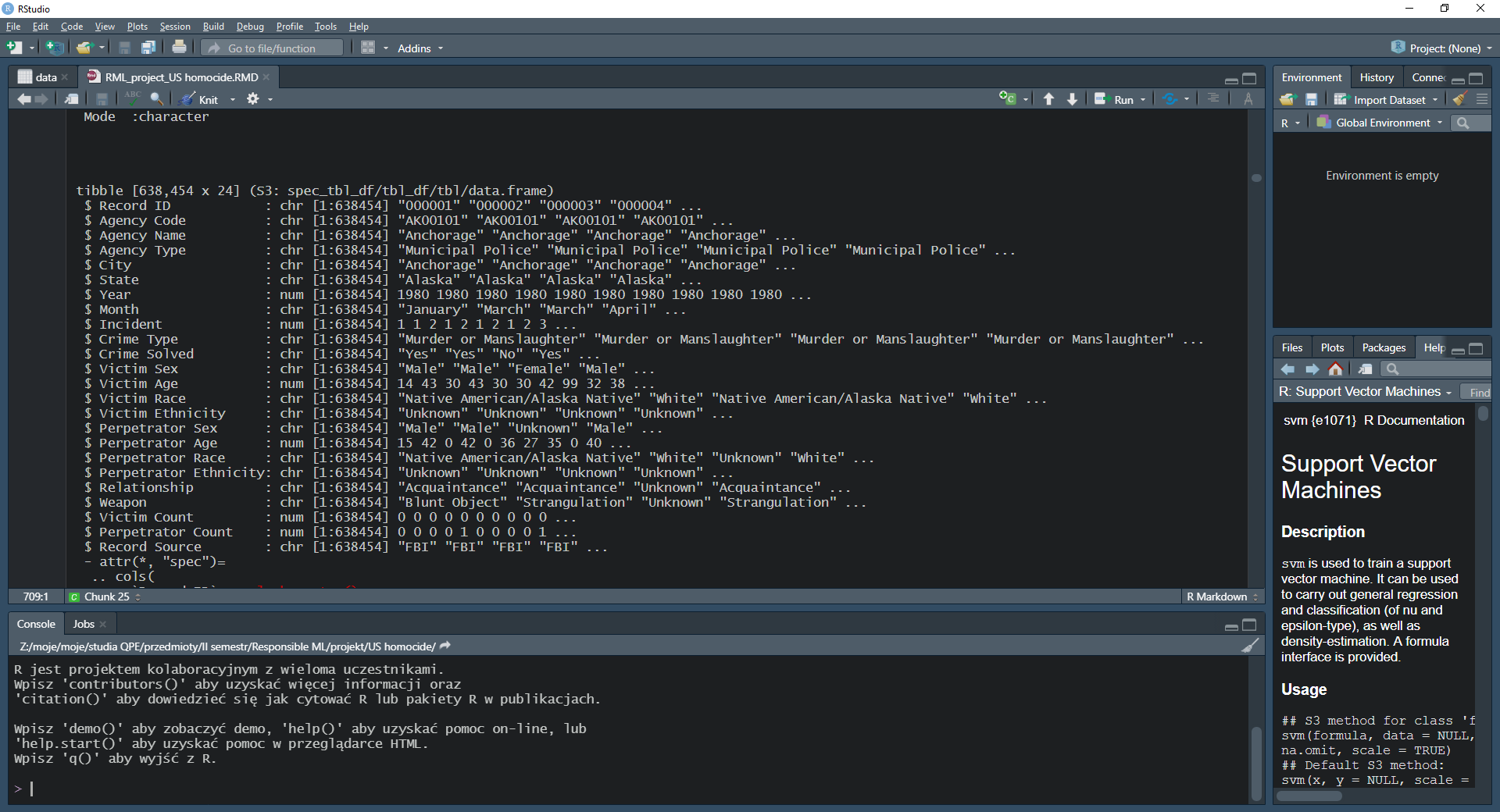
**Problem Context**

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 and Freedom of Information Act data on more than 22,000 homicides that were not reported to the Justice Department. This dataset includes the age, race, sex, ethnicity of victims and perpetrators, in addition to the relationship between the victim and perpetrator and weapon used.

**Data**

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

The data set 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.



**Target variable**

* Crime Solved

**Protected variables**

* Victim\_Sex
* Victim\_Race
* Victim\_Age

**Objectives**

The goal of the analysis is to determine whether fairness of the model is maintained. To that end, we analyze whether identified protected variables affect the outcome of the crime investigation. In the data set victim's age, race and sex were identified as protected variables and whether a crime was solved or not was identified as target variable. In addition, explainability analysis was also conducted.

To sum up, the following are the objectives of the project:

1) Exploratory Data Analysis

2) Construction of predictive model

3) Explainability analysis

4) Assessment of model's fairness

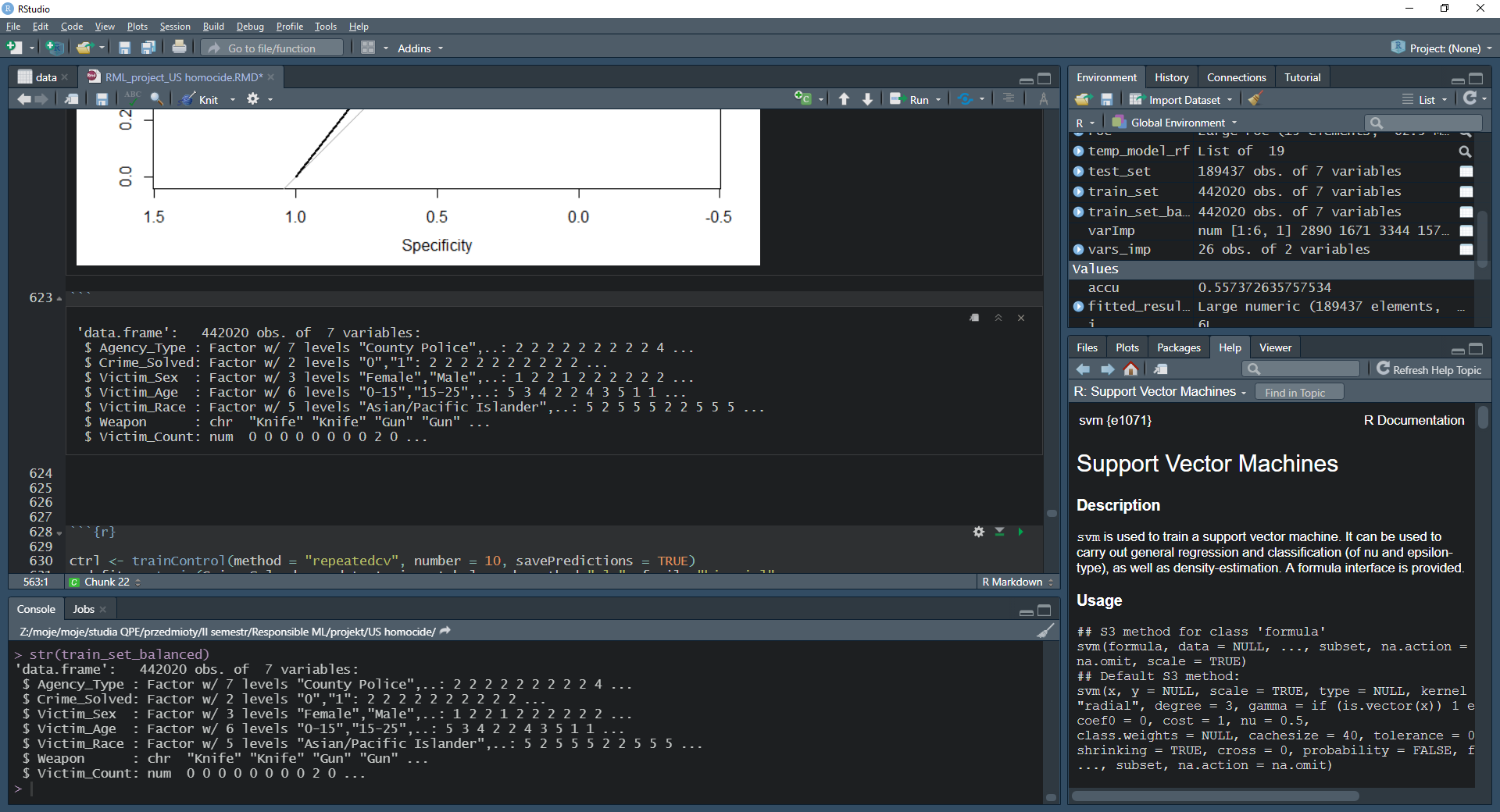
**In an essence, the goal is to see if depending on victim’s race, victim’s sex or victim’s age crimes were solved more or less often than general average.**

**ML algorithms and variables**

I plan to use two machine learning algorithms:

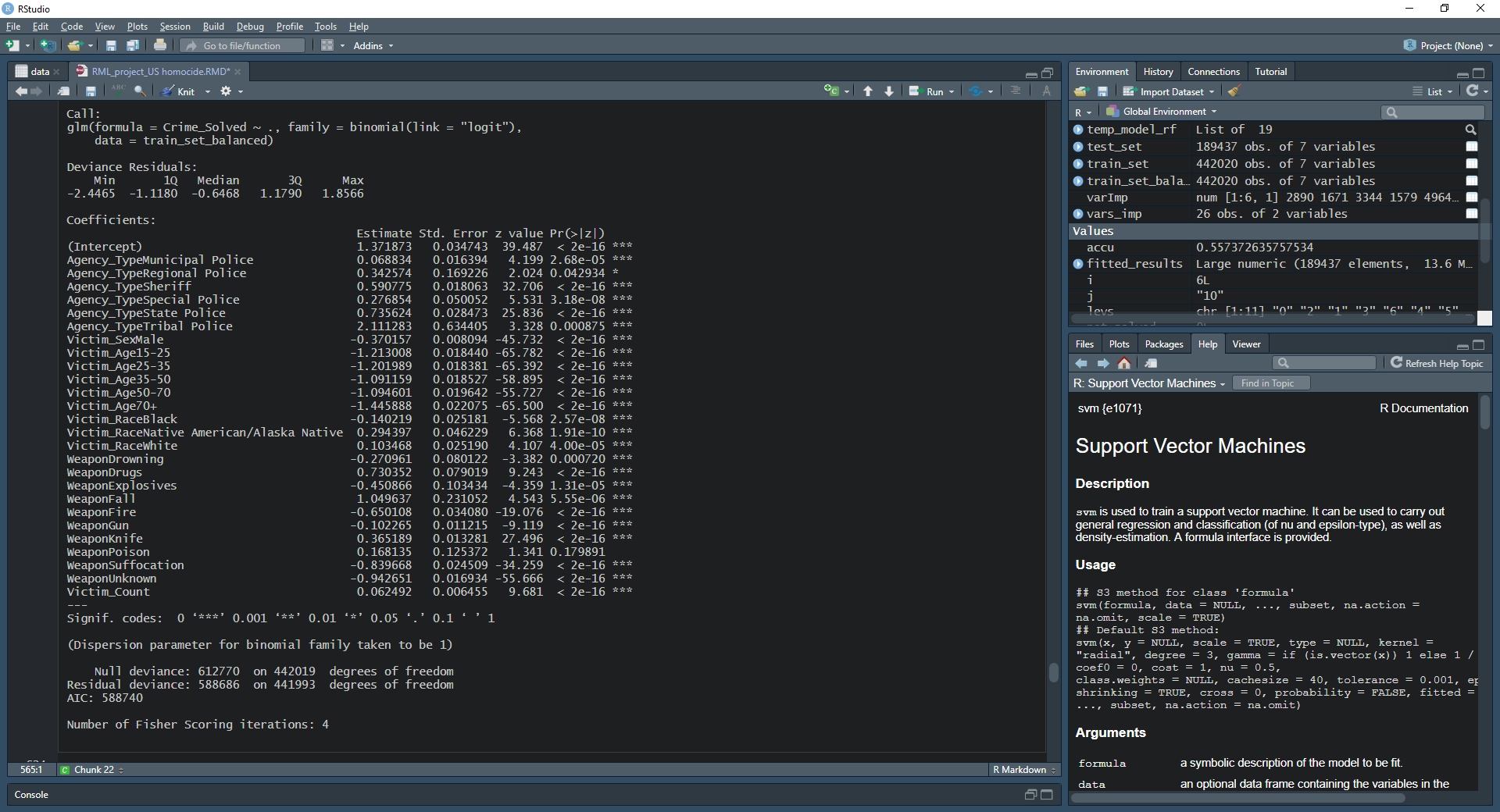
1. LogisticRegression
2. Random Forest

For each of the above I used the following dataset. Crime\_Solved is a target variable and the remaining 6 variables are independent variables.

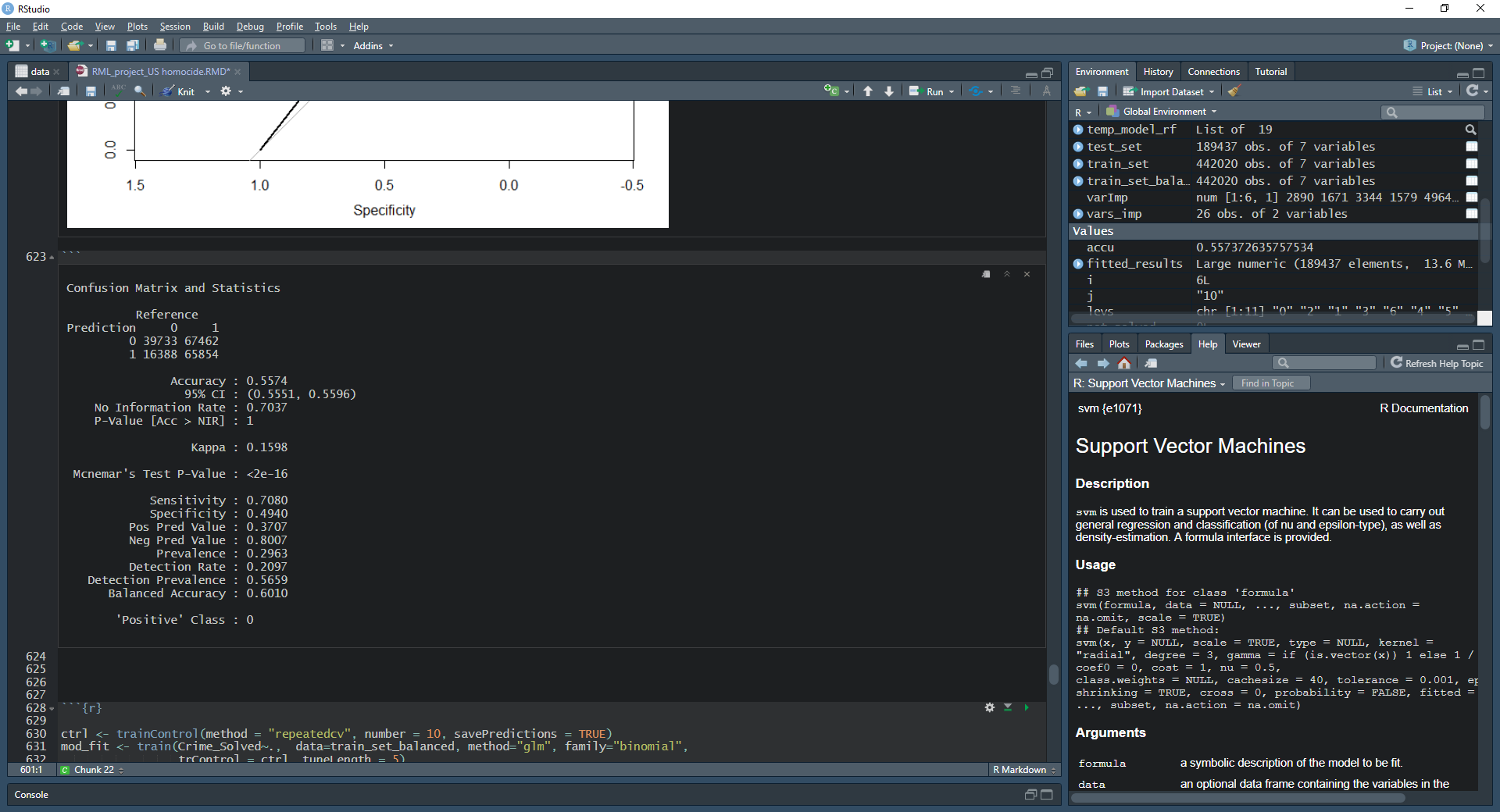


Below I present some initial results.

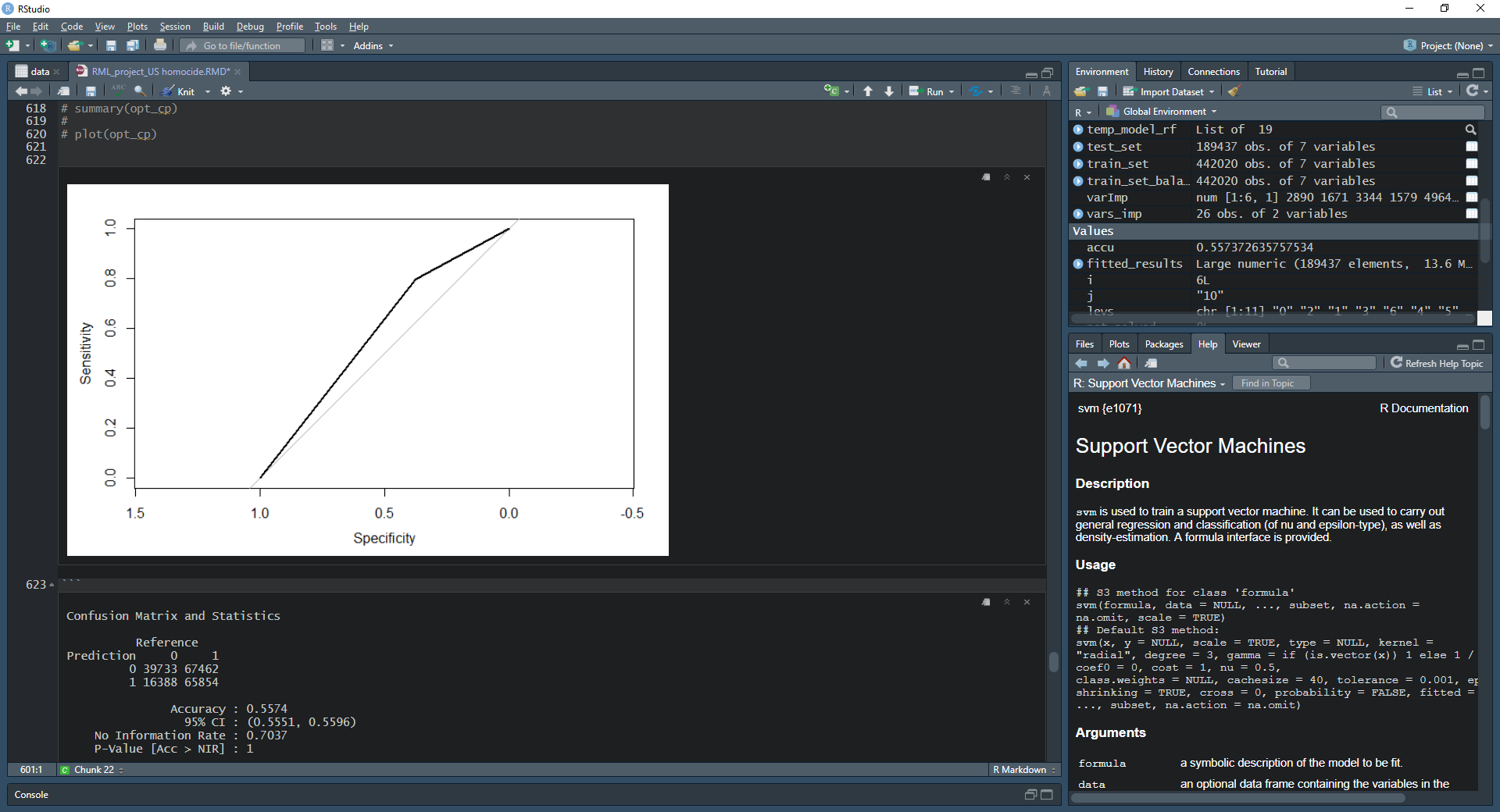
1. Logistic regression



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.

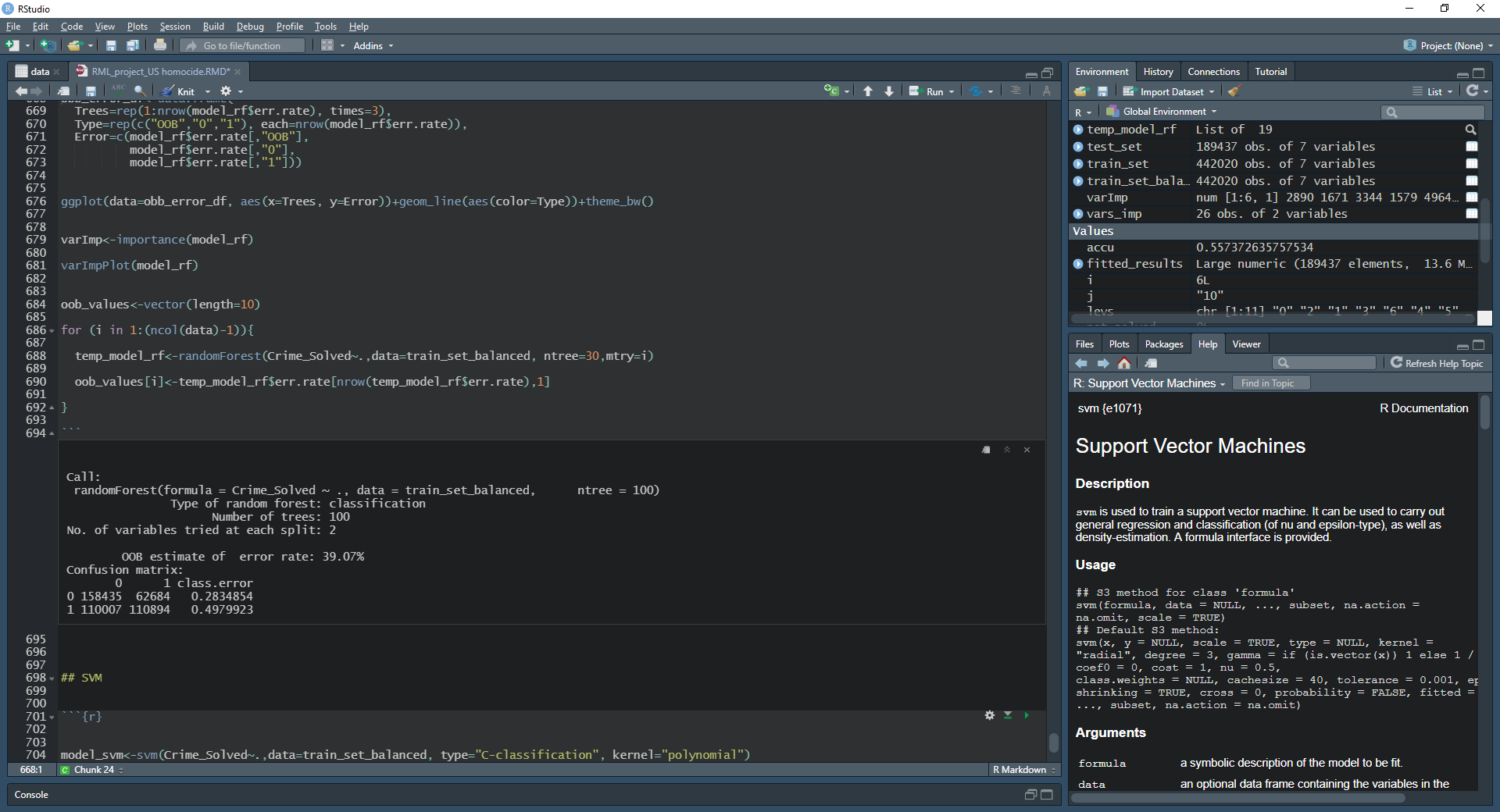


Accuracy: 0.5857

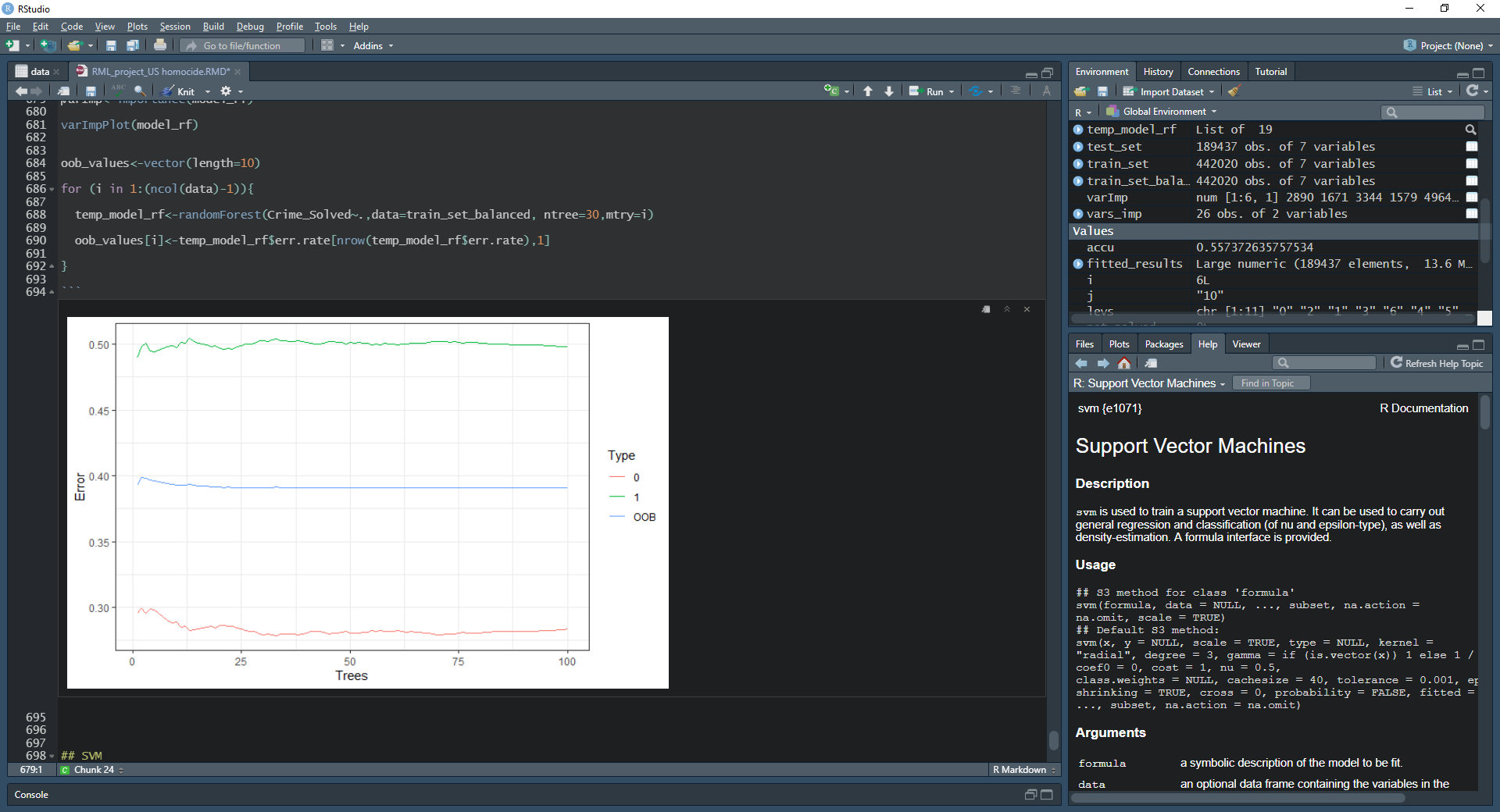
Clearly, the results are poor. I tried to improve them by running 10-fold cross validation on the training set, but the results were the same as with the simple logistic regression.

1. **Random Forest**

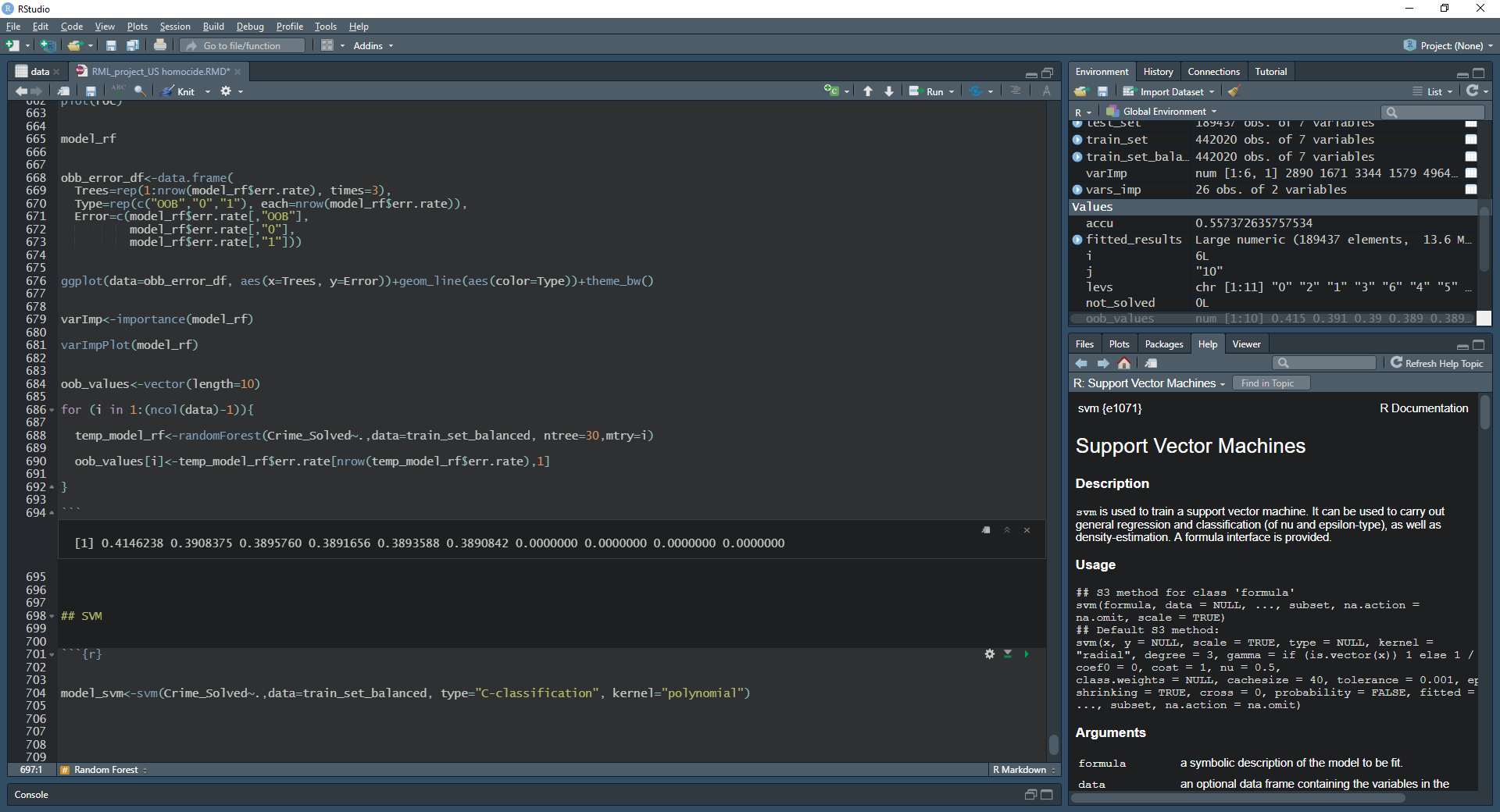
I have run the following analysis**.**



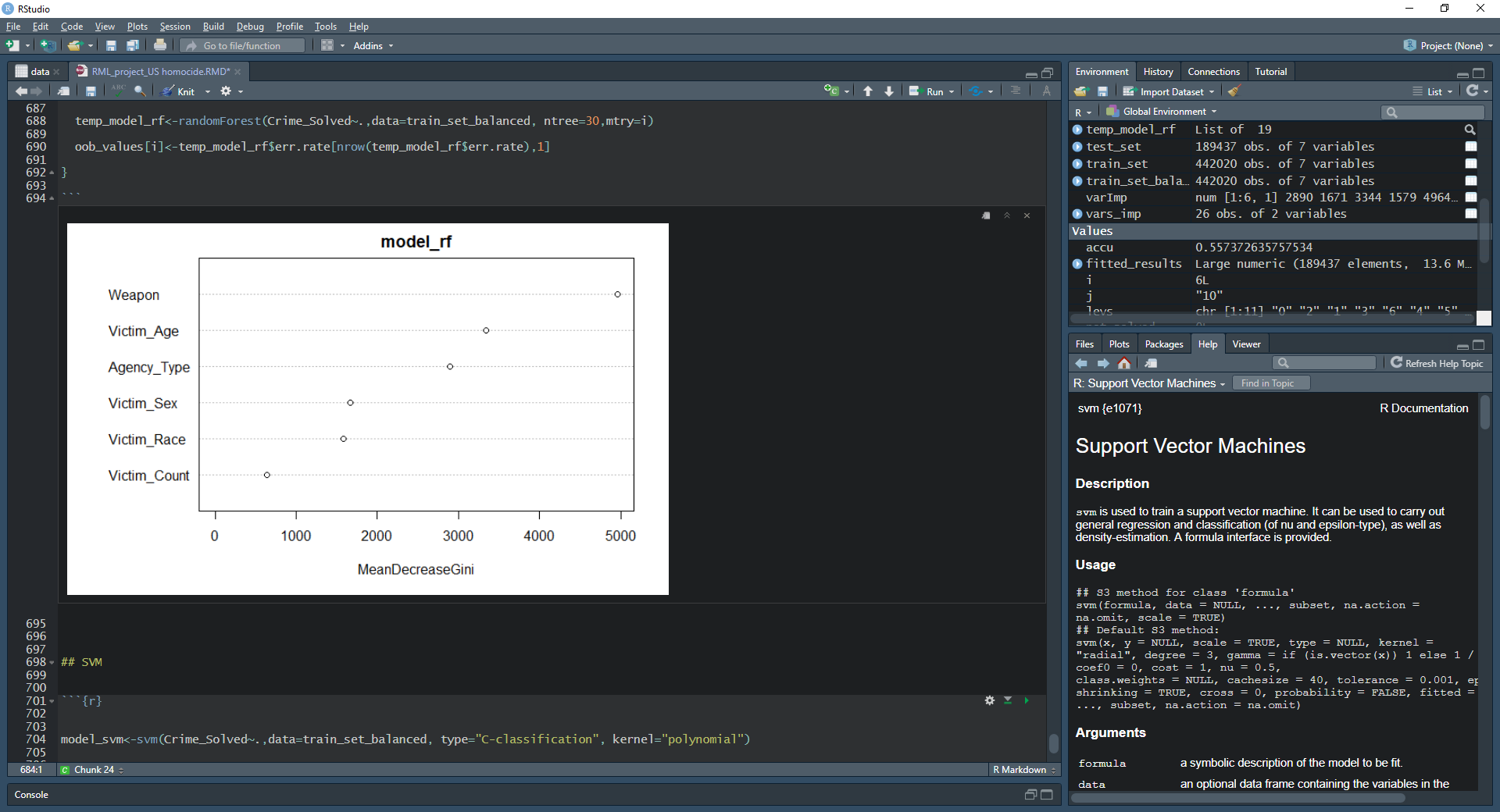
I used 100 trees in the random forest formula, as error rates are stable then.



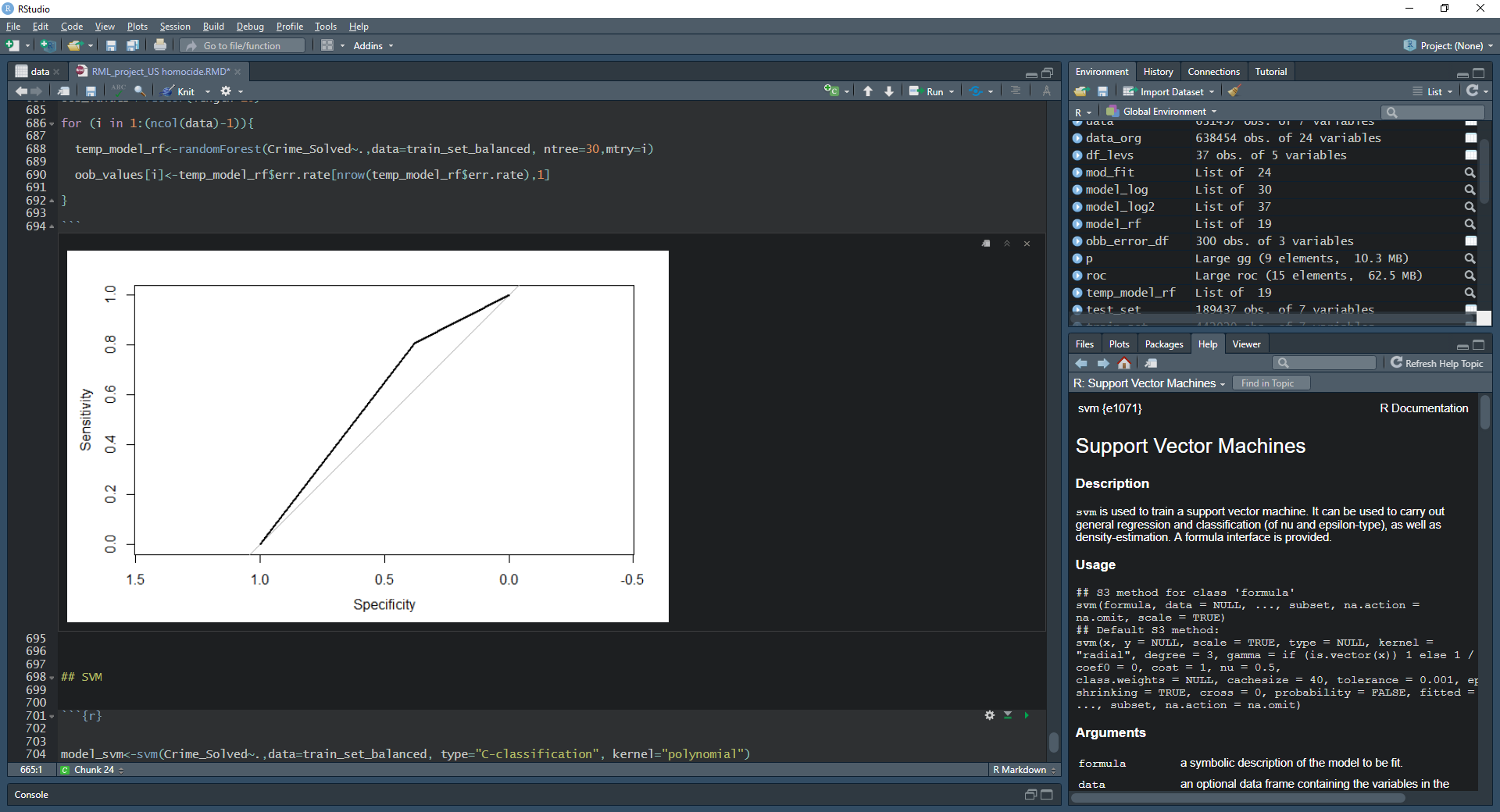
Out-of-bag error depending on the number of variables evaluated at each step of random forest (from 1 variable to max 6 variables). Error rate does not improve almost at all after 2 variables.



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