Study Of Consumer Complaints Dataset To Identify Whether Consumer Will Be Disputed Or Not

Data Mining Final Project
Ashwini Pisal, Masters Student

Georgia State University

Abstract

The Consumer Financial Protection Bureau ("CFPB" or "Bureau") was established under Title X of the Dodd-Frank Wall Street Reform and Consumer Protection Act ("Dodd-Frank Act"). To create a single point of accountability in the federal government for consumer financial protection, the Dodd-Frank Act consolidated many of the consumer financial protection authorities previously shared by seven federal agencies into the CFPB and provided the Bureau with additional authorities to:

- Conduct rulemaking, supervision and enforcement with respect to the Federal consumer financial laws.
- Handle consumer complaints and inquiries.
- Promote financial education.
- Research consumer behaviour.
- Monitor financial markets for risks to consumers.

Consumer complaints dataset is maintained by CFPB to achieve above targets.

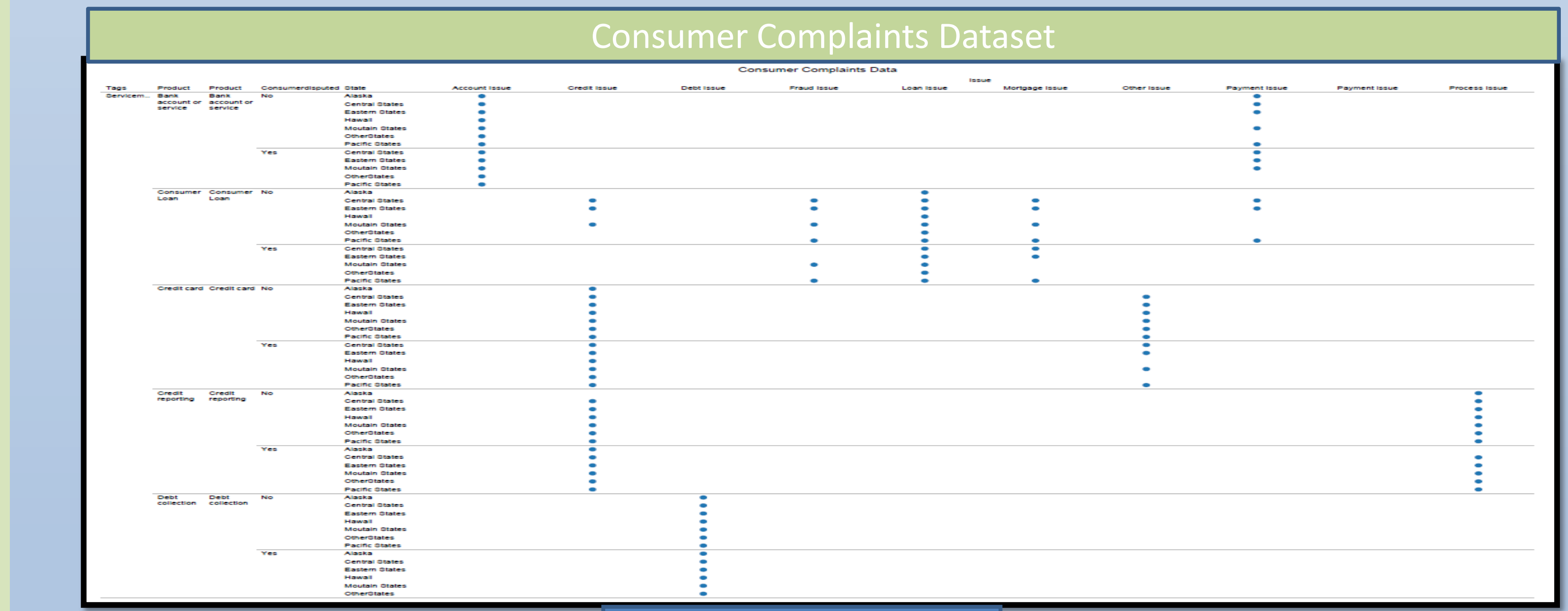
Introduction

Consumer Complaint Database taken from www.datagov.com . This dataset contains complaints registered by financial product consumers. It has various attributes such as financial products, issue description, company, states, consumer disputed, etc. The CFPB does not verify the accuracy of all facts alleged in the complaints, but takes steps to confirm a commercial relationship between the consumer and the identified company exists.
For study I have considered only five attributes Product, Company, States, Issue, Tags and Consumer Disputed. To identify whether consumer will be disputed or not?

Model Used For Prediction

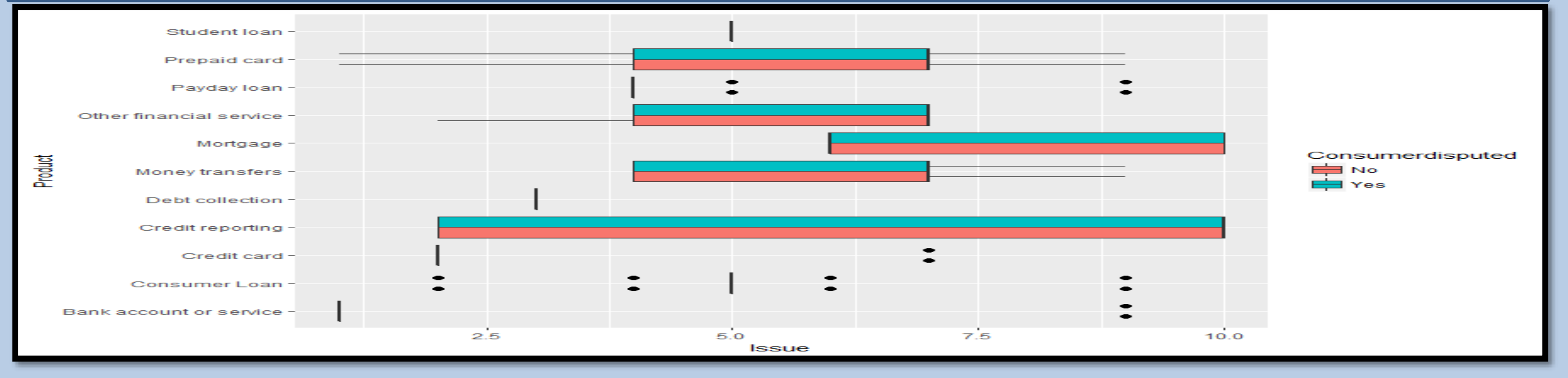
I have used classification model to predict whether consumer will be disputed or not. My study includes implementation of Decision, Regression trees (using rpart and Random Forest packages) and Naïve Bayes Classifier. To work with this Dataset , it is divided into training and test data. Here onwards will start with our process

Consumer Complaints Dataset

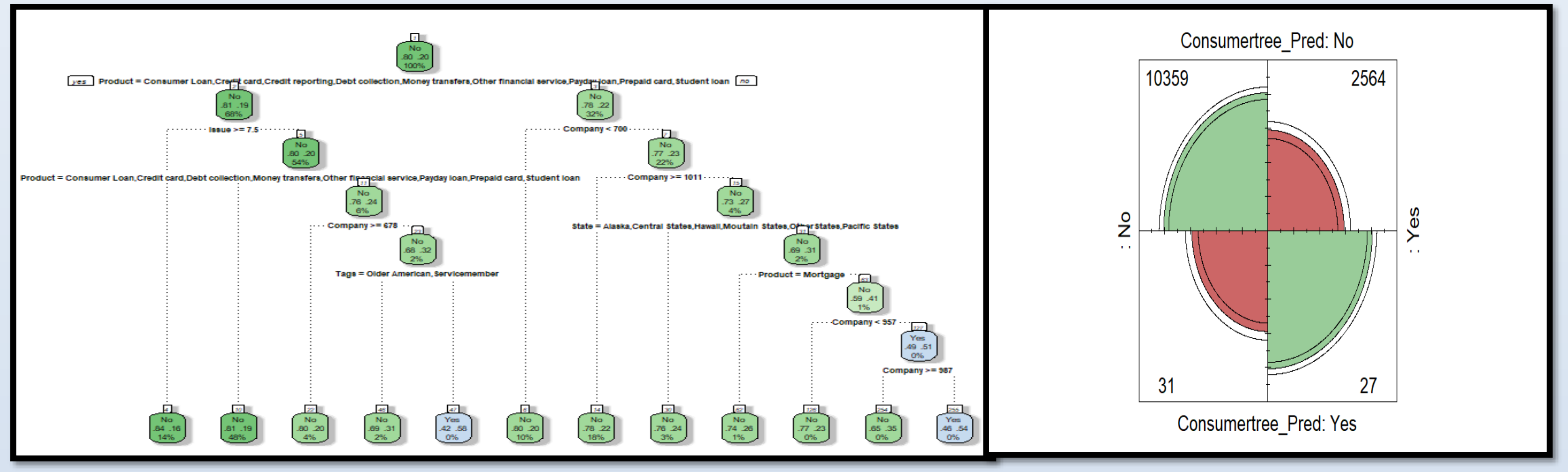

Consumer Complaints Data

This dataset is very big for analysis and better Prediction it is divided into Training and Testing

Consumer Complaints Training set Box Plot



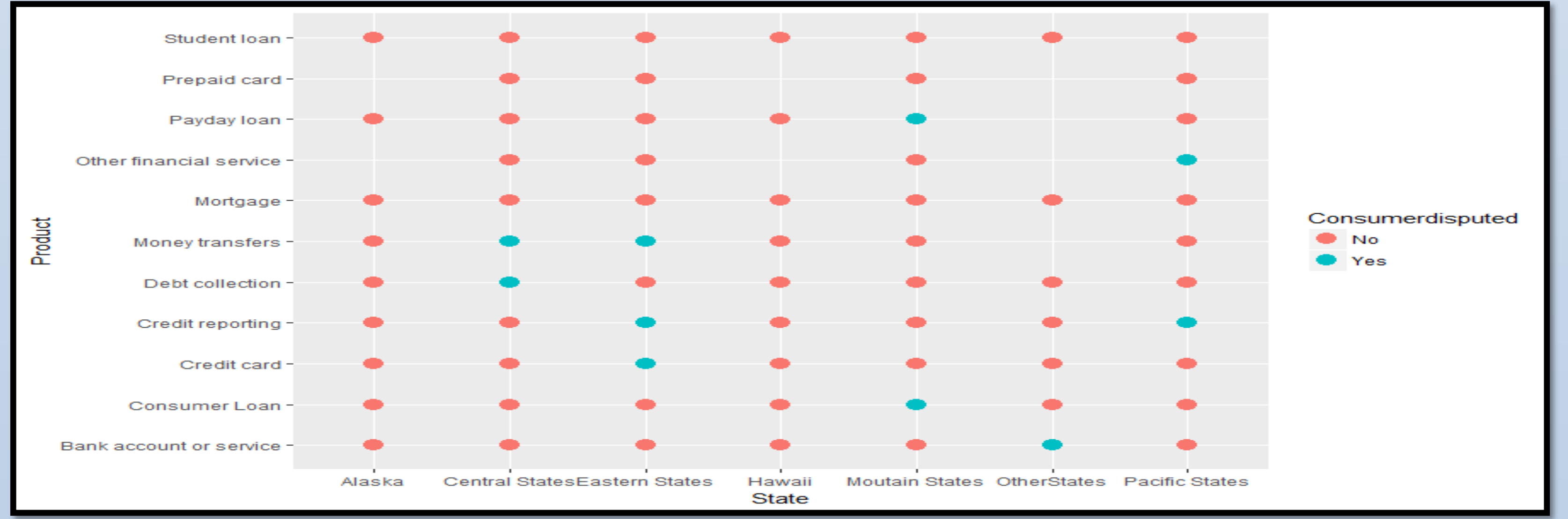
Decision Tree (Rpart) for Training Dataset



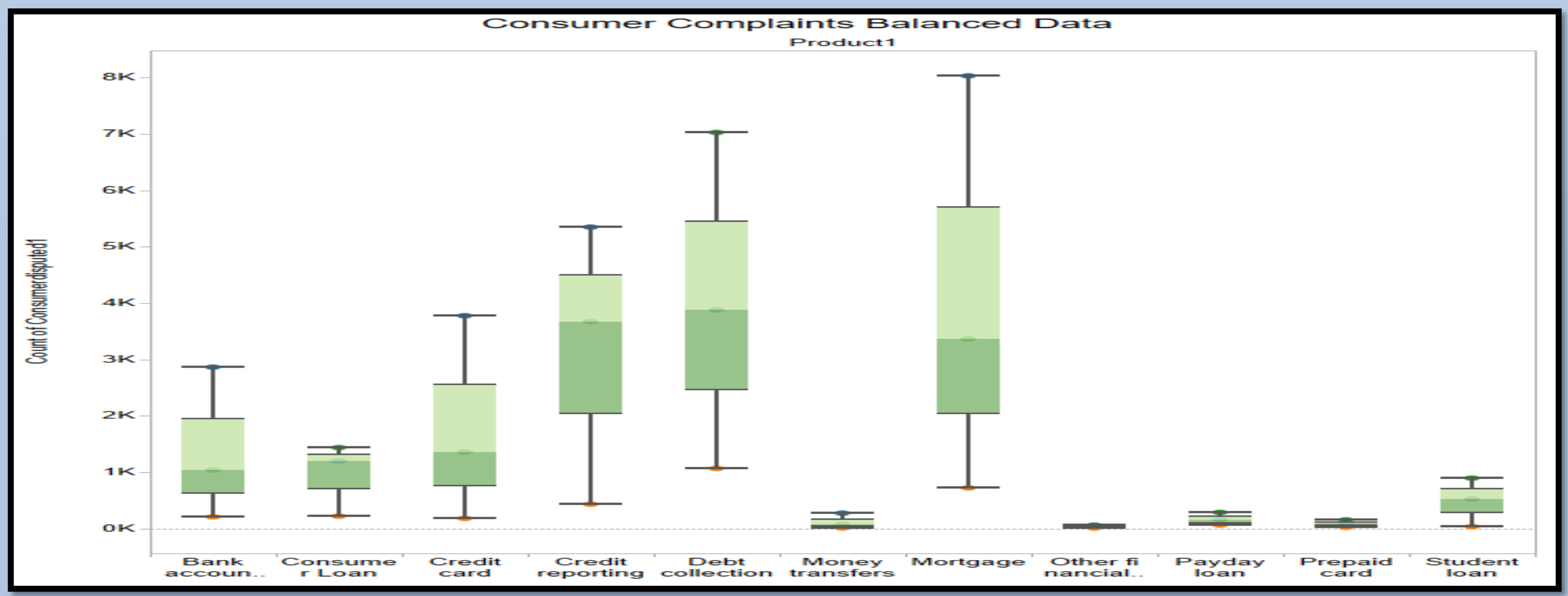
Result Analysis and Issues

After analysing the few results of Decision Trees, I have found that given dataset is imbalanced. It has 80% No results for consumer dispute and 20% result disputed. To resolve this issue I have used concept of under sampling, Over sampling and Both sampling techniques using ROSE Package.

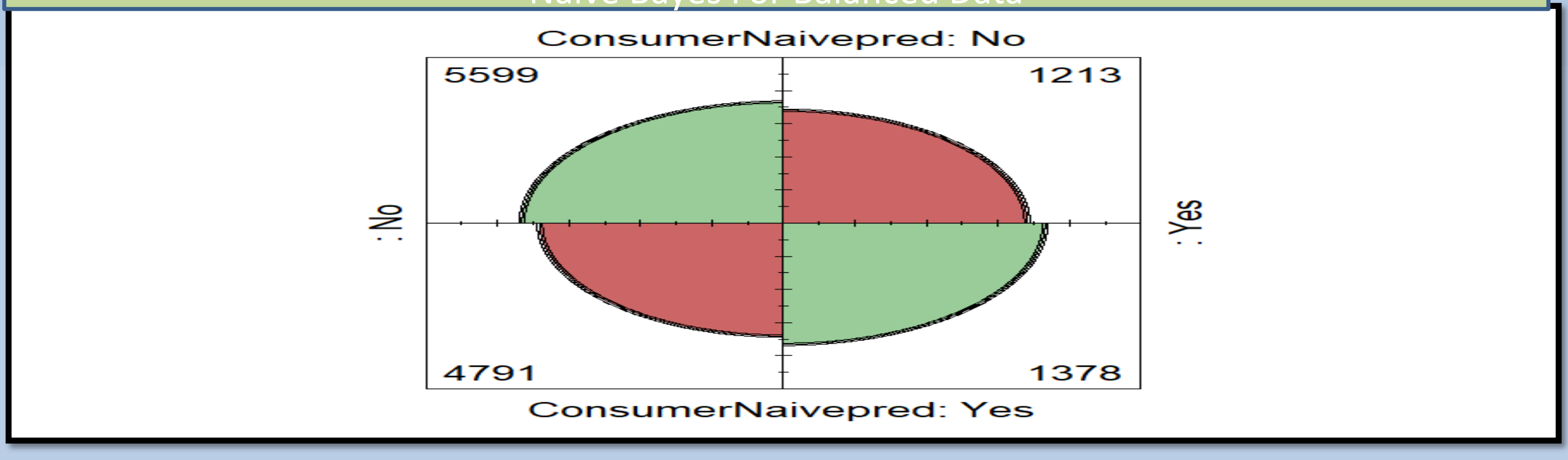
Consumer Disputed attribute of Original Data



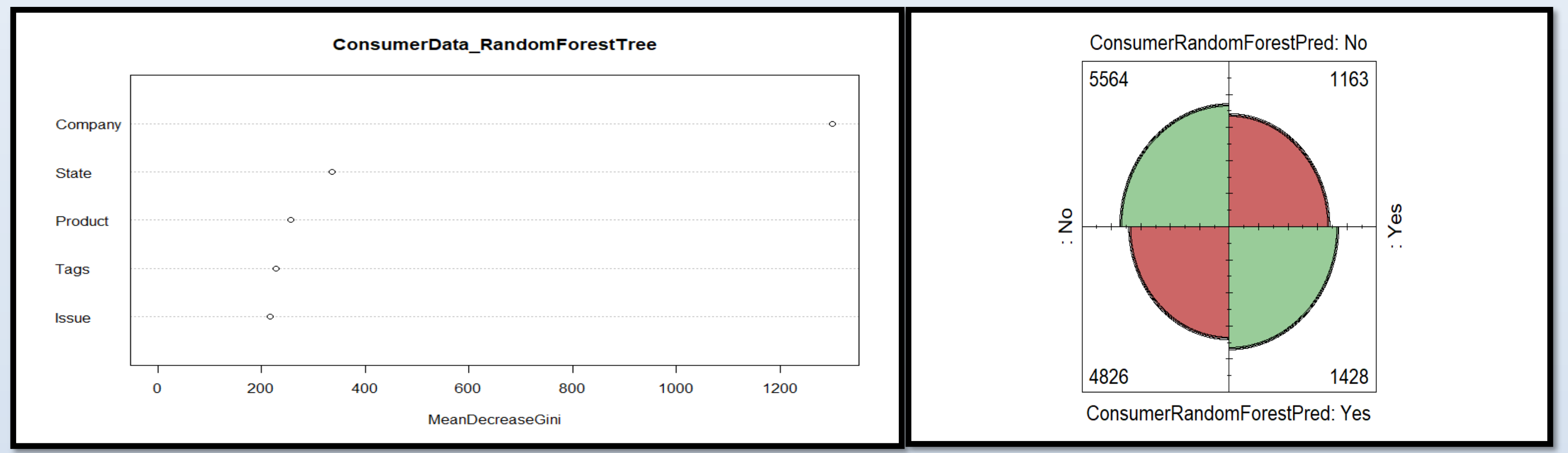
Data is Sampled using ROSE Package In R Box plot for balanced data



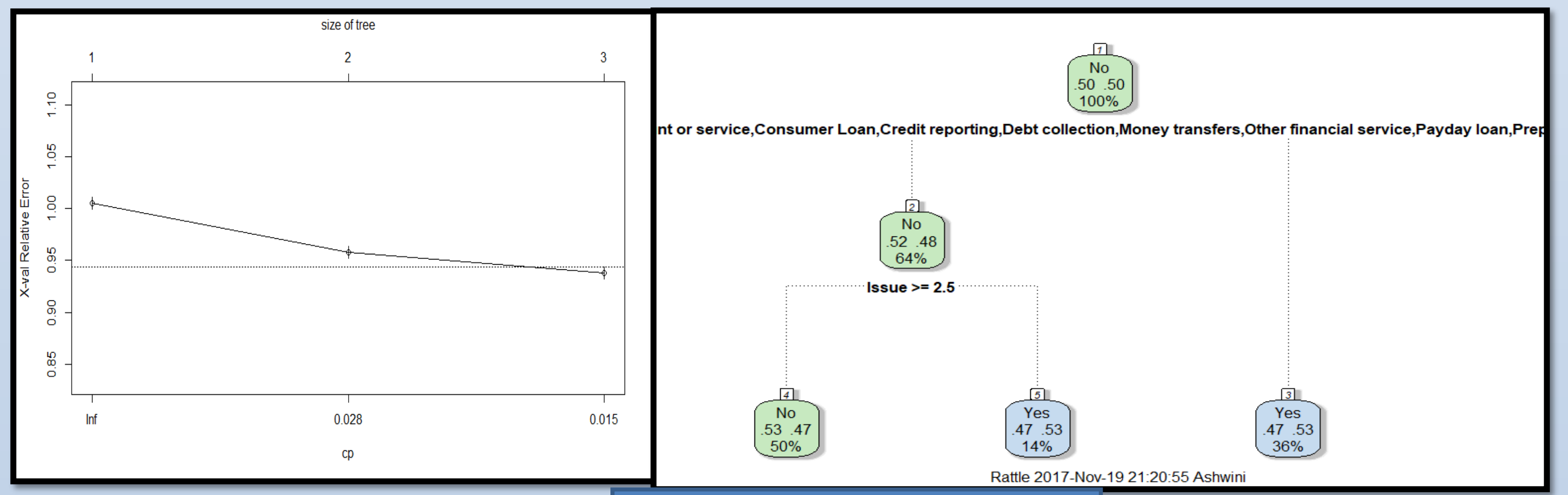
Naïve Bayes For Balanced Data



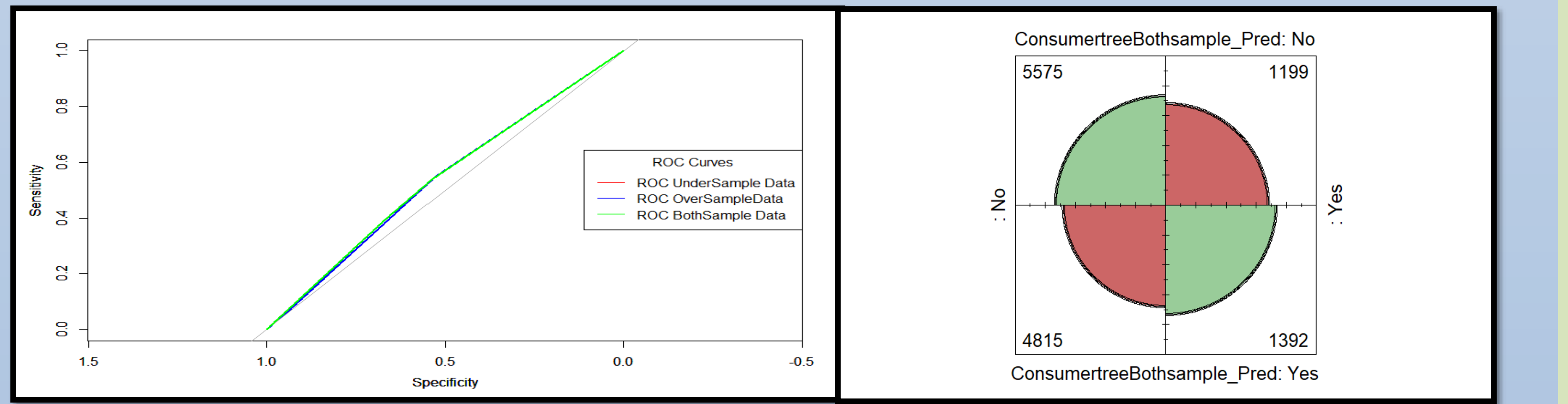
Results of Balanced data for Random Forest



Results of Balanced data after applying both sampling technique for Decision Tree



Results for Under ,Over and Both Sampled datasets



Original & Both Sample Data Accuracy

Original Data
Accuracy : 0.8001
Sensitivity : 0.99702
Specificity : 0.01042

Both Sample Data
Accuracy : 0.5367
Sensitivity : 0.5366
Specificity : 0.5372

Under ,Over Sample Data
Accuracy : 0.5311
Sensitivity :0.5270
Specificity : 0.5473

Conclusions

On analyzing the results for original data and other three sampled data, we can say that Dataset sampled with under and over technique is better than other dataset results. Though it has less accuracy compared to original data but it has good sensitivity and specificity.
All these classifiers have predicted the label "NO". Based on this information we can say