Prediction of Consumer Disputes about Financial Complaints Response

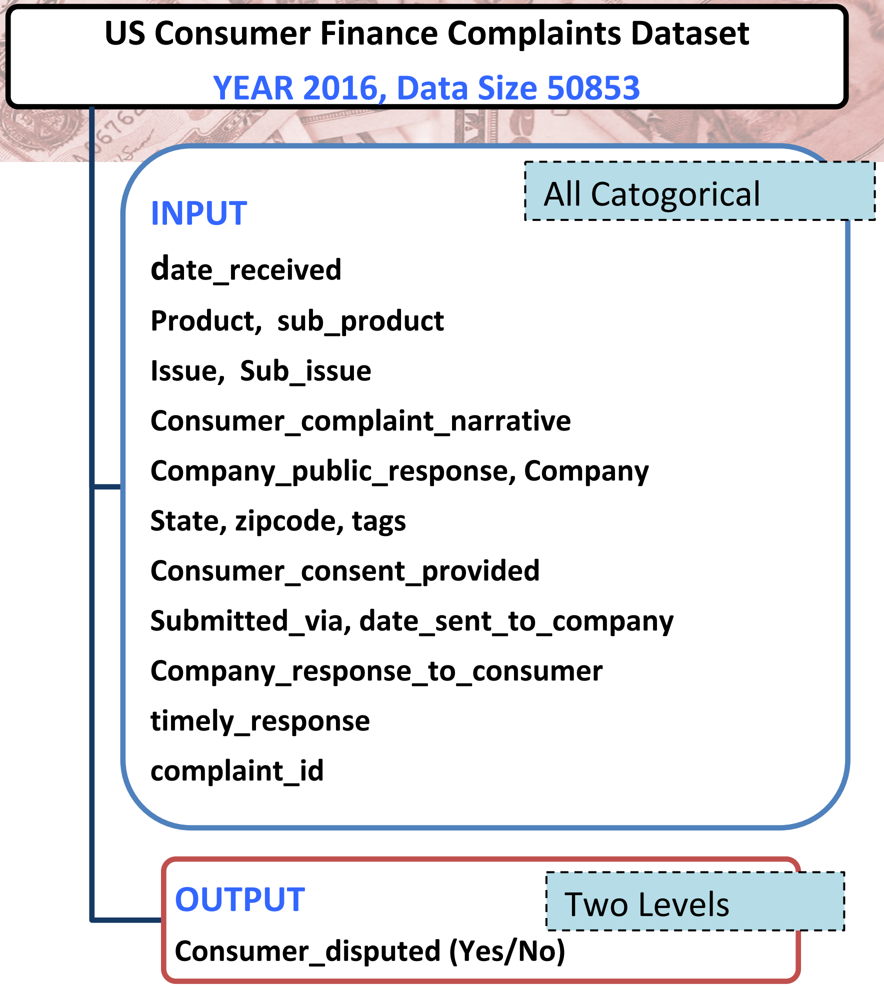
Team 2 Members:
Fengmei Liu, Xiaoyan Chong and Weiqian Hou

Outline

- Introduction and motivation
- System Design & Implementation details
- Experiments of concept evaluation
  - Data Preprocessing
  - Logistic regression
  - Naive Bayes
  - Gradient Boosting Tree
- Conclusion and Discussion

Problem to Solve

- **Data**
  - Kaggle dataset “US Consumer Finance Complaints”.
- **Background**
  - Consumer Financial Protection Bureau (CFPB) sends consumers’ complaints about financial products/services to companies for response
- **Objective**
  - Classification: Consumer Disputed(YES/NO) from a knowledge of complaint patterns.
- **Meaning**
  - Use as a reference for companies to understand their customer service
  - Guide customers to follow the correct way of feedback to get complains successively solved



US Consumer Finance Complaints Dataset

YEAR 2016, Data Size 50853

INPUT

All Catogorical

date_received

Product, sub_product

Issue, Sub_issue

Consumer_complaint_narrative

Company_public_response, Company

State, zipcode, tags

Consumer_consent_provided

Submitted_via, date_sent_to_company

Company_response_to_consumer

timely_response

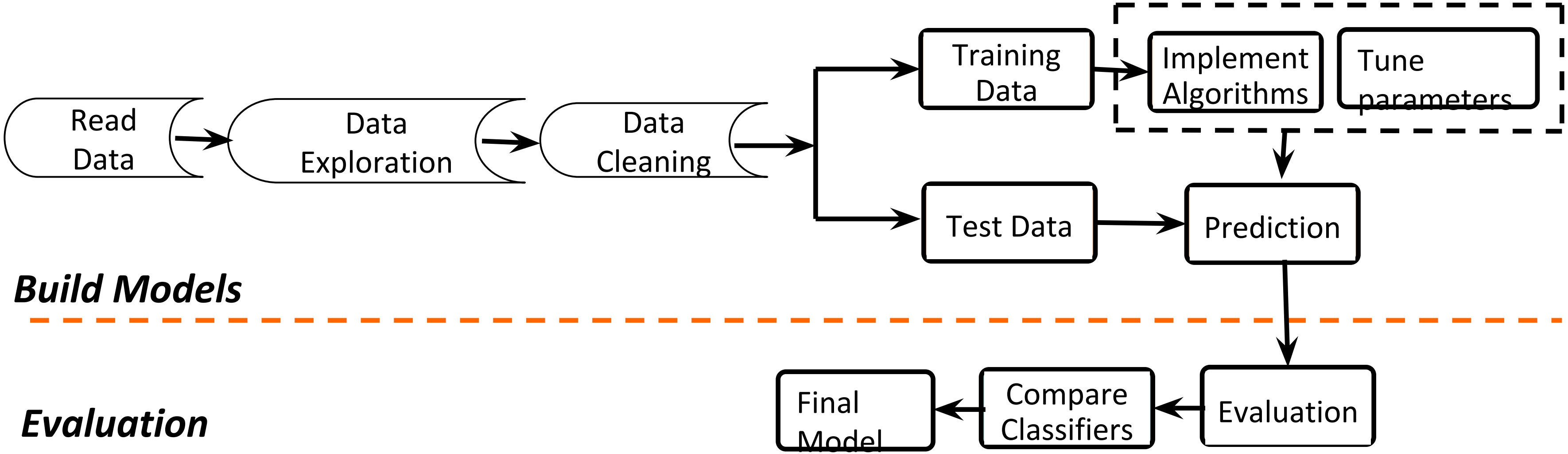
complaint_id

OUTPUT

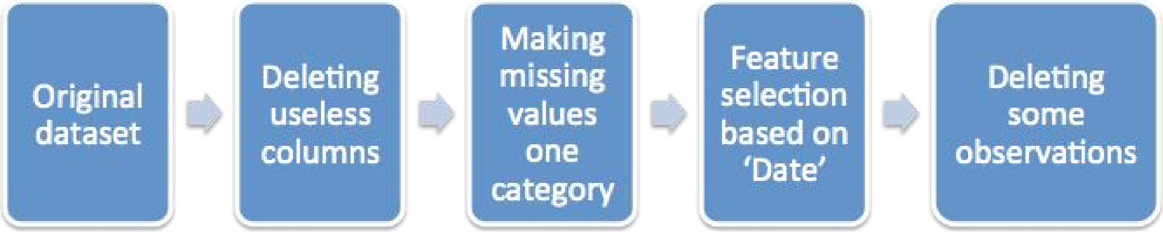
Two Levels

Consumer_disputed (Yes/No)

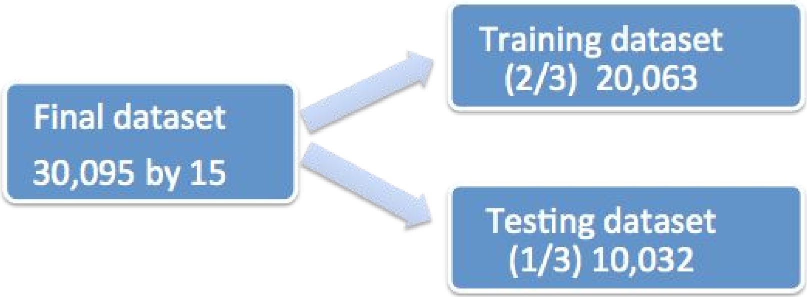
System Design and Process Flow



Data Preprocessing



Original dataset → Deleting useless columns → Making missing values one category → Feature selection based on 'Date' → Deleting some observations



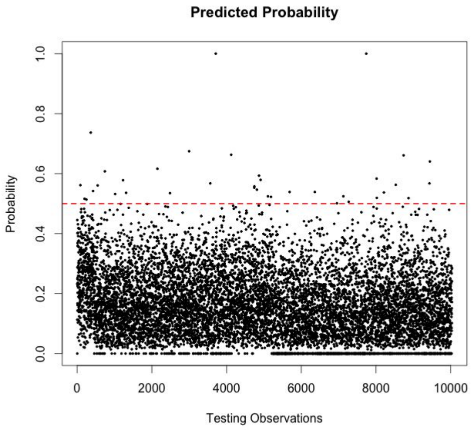
Final dataset

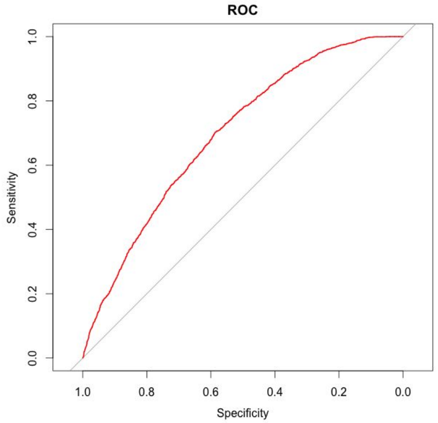
30,095 by 15

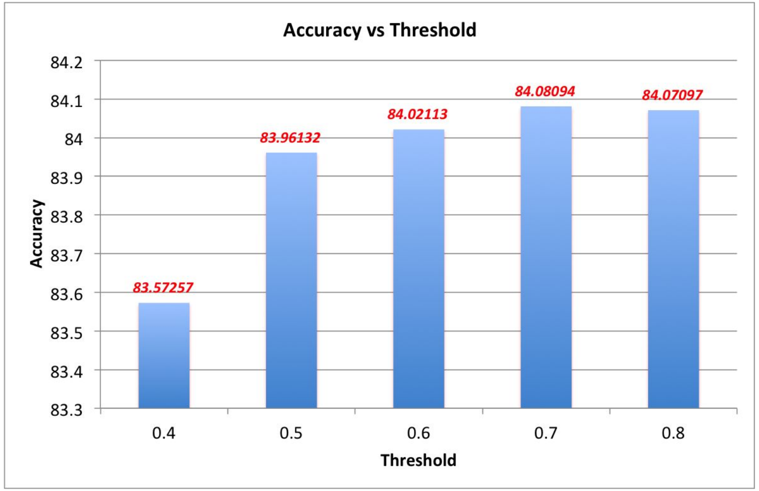
Training dataset
(2/3) 20,063

Testing dataset
(1/3) 10,032

Logistic Regression







AREA under the curve is: 0.6963

Reason?

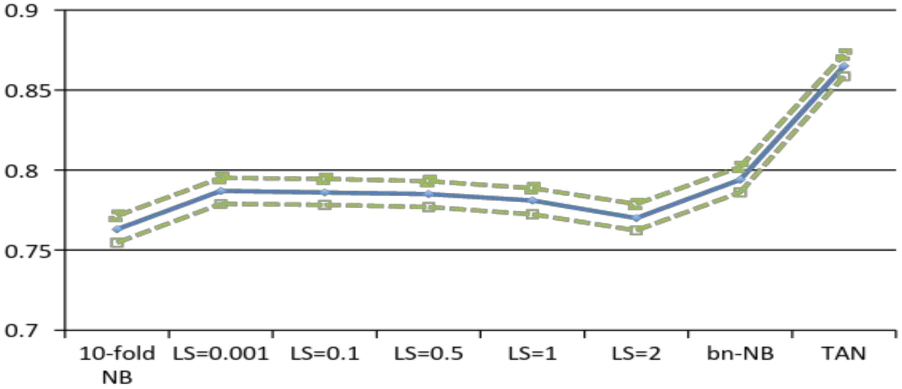
Logistic Regression

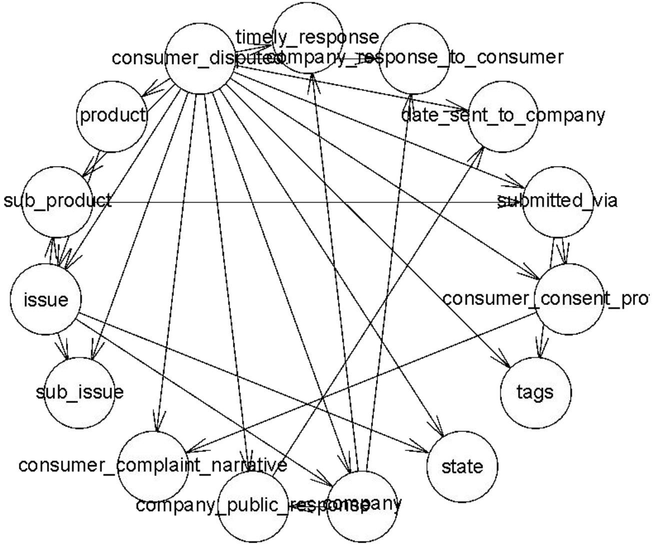
- **Pros:**
  - fast and intrinsically simple
  - low variance and so is less prone to overfitting

- **Cons:**
  - Doesn’t handle large number of categorical variables well
  - Does not give us much space to improve the final predicting accuracy.

Naïve Bayes & Tree Augmented Naïve Bayes(TAN)

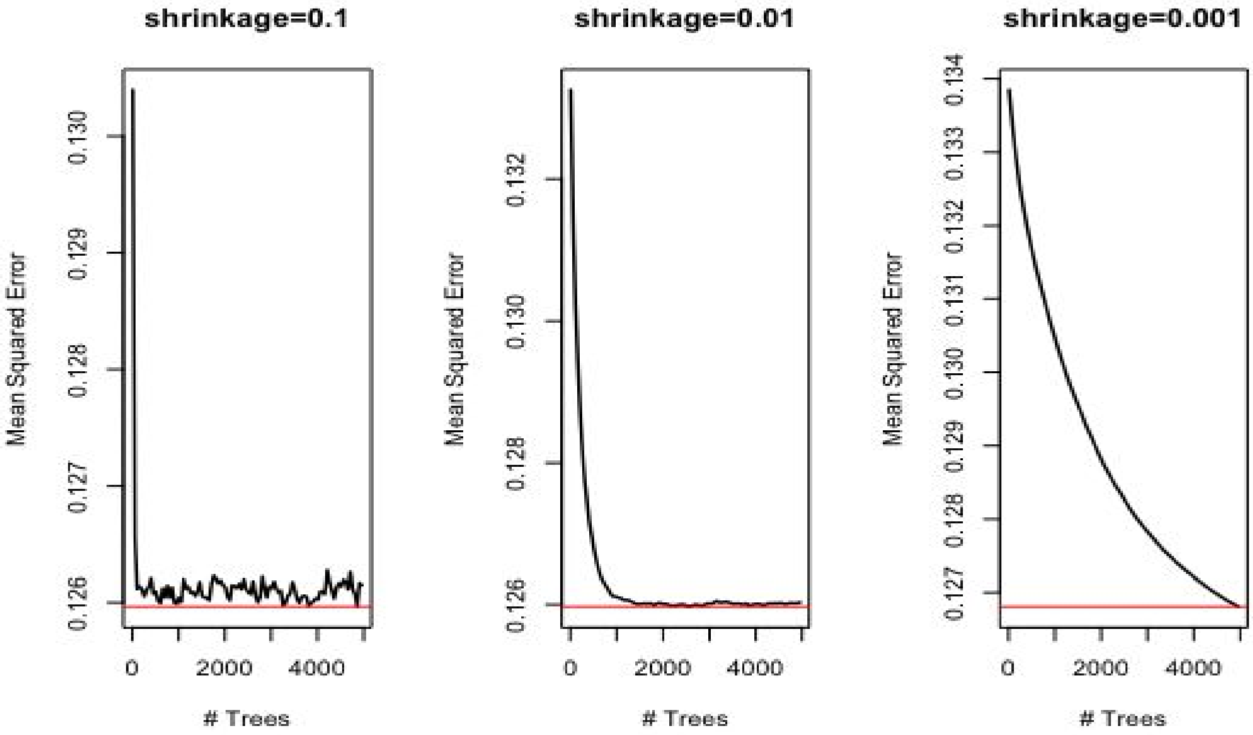
- **Models**
  - Simple Naïve Bayes, Cross Validation 10-fold
  - Laplace Smoothing (0.001, 0.1, 0.5, 1,2), Cross Validation 10-fold
  - TAN
- **Result**
  - TAN has the best accuracy **0.865** and Narrowest 95% **C.I**.
  - Laplace smoothing has no improve as the test data has no new categories.



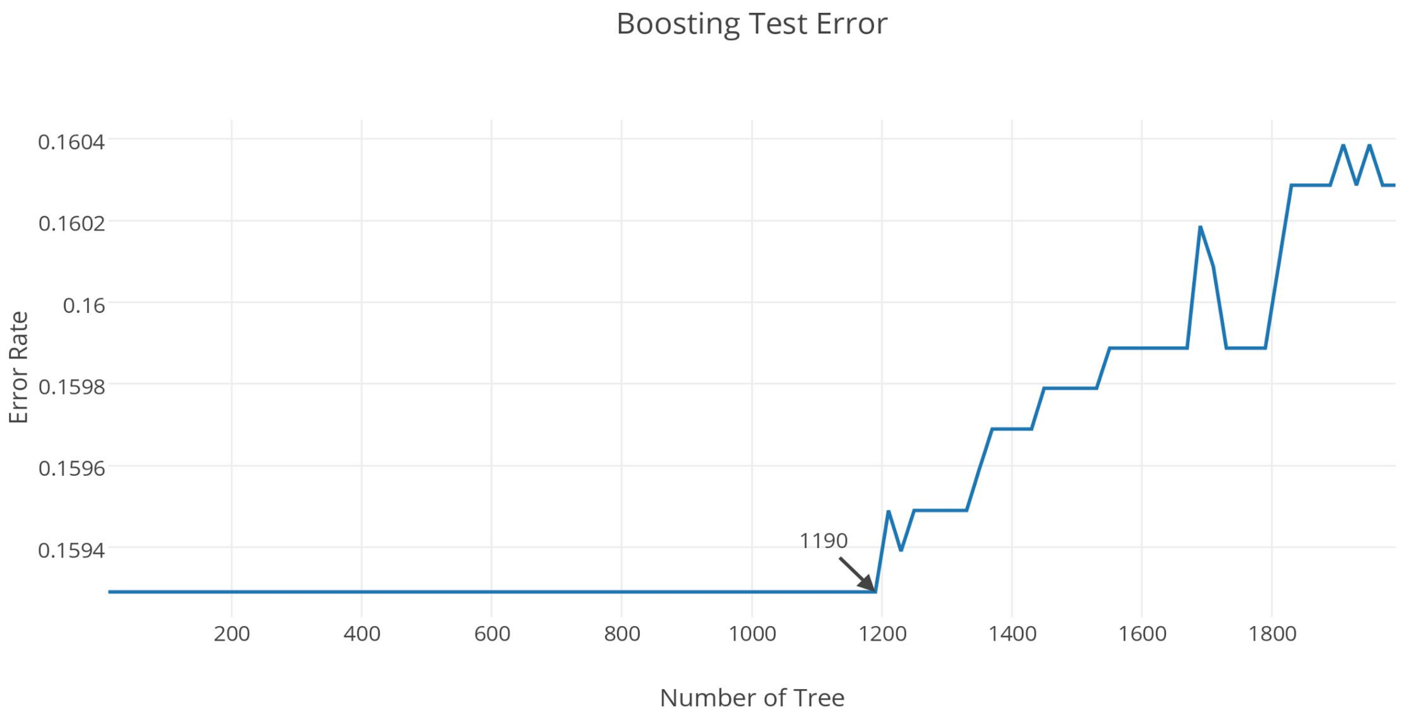


Run Result: TAN model shows the correlation between inputs and output

Decision Tree (Gradient Boosting)



Decision Tree (Gradient Boosting)

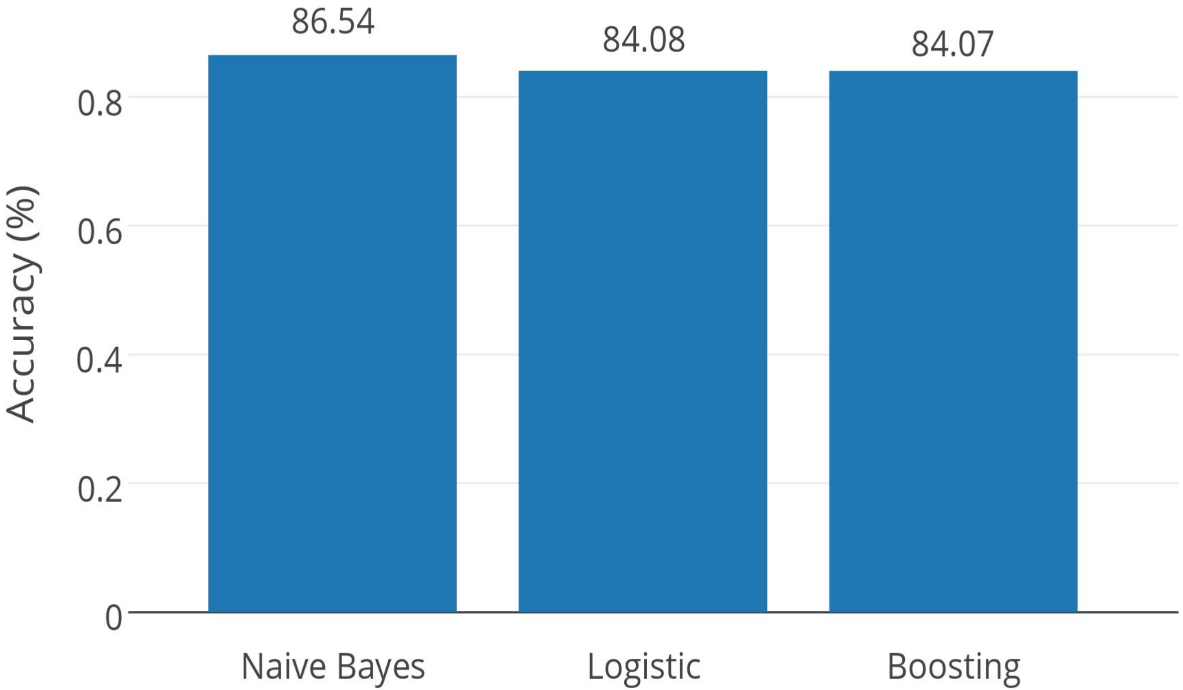


Gradient Boosting Tree

- **Pros:**
  - Yields a very accurate classifier.
  - Faster than other boosting methods such as adaptive boosting.

- **Cons:**
  - Hard to tune the model because of these parameters.
  - Easy to get overfitting.
  - Not very speedy.

Comparison of Three Algorithms



Thank you