

Complaint Categorization

...

Caleb Christensen and RJ Cass
Stat 536

Table of Contents

1. Introduction/Abstract
2. Feature Construction
3. Model Review
4. Research Questions
5. Suggested Next Steps

Abstract

Resolving complaints in a timely manner begins with sending each complaint that comes in to the appropriate department. We have around 125,000 complaints in a dataset paired with the corresponding department the complaint was sent to. We train a **random forest** on these data and obtain a classification rate of **78.9% In-Sample**, and **78.4% Out-of-Sample**. Using the model, we determine what components of a complaint are the most useful for categorization, identify which departments are commonly miscategorized, and predict which department the new, unresolved complaints should be sent to.

Background

How quickly a company is able to respond accurately to a customer's complaint has a large impact on the client's opinion of that company.

We have a dataset of complaints filed to companies and which department the complaint was sent to for help. Using this dataset, we want to create a model to help us determine which department a complaint should get sent to to quickly answer the client's concerns.

Questions of Interest

1. How accurately are you able to classify complaints?
2. If possible, what key words or symbols or explanatory variables are useful in classifying complaints?
3. Are there departments that are commonly confused? That is, are complaints commonly classified as going to department “A” when it should go to department “B”?
4. Of the provided sample complaints, what department(s) does your model think they should be directed to?

Data

The base dataset includes only the complaint text, and the department the complaint was sent to. To create a model on this data, we first had to create features from the complaint text.

We created a total of 41 features. Some examples are:

- Avg. Word Length
- Word Count
- Presence of Keywords
- Count of words associated with each department

Features

To develop features, we explored online what some of the most common open-text features were. Among them we identified:

- Word-based features
- Letter-based features
- Sentence counts

Features - Word-based

To develop our word count features, we

1. Identified the most commonly used words for complaints in each department, then developed word 'profiles' for departments (common words that appeared to go together)
2. Created counts of common words used in multiple departments
 - a. We attempted basic lemmatization (using the root of words, ie. run == running) by trying to include all variations of a word in our search criteria
3. Counted the overall, unique, and repeat words in each complaint

Features - Letter/Sentence Based

Letter Based:

- We used the ratio of capital letters (ie. $\frac{\# \text{ Capital Letters}}{\text{Total Letters}}$)

Sentence Based:

- We got the total count of sentences per complaint

Model 1

The first model we tested was a Random Forest. Our random forest had 1000 total trees and used the gini impurity index for splitting branches.



Strengths

- Allows us to prediction/classification
- Can calculate classification rates.
- Can easily get variable importance
- No basic assumptions besides assuming that the data are 'good'

Weaknesses

- Can interpret, but more complex than more simple models

Model 2

We tested a variety of levels, numbers of nodes, and activation functions. Decided to use model with:

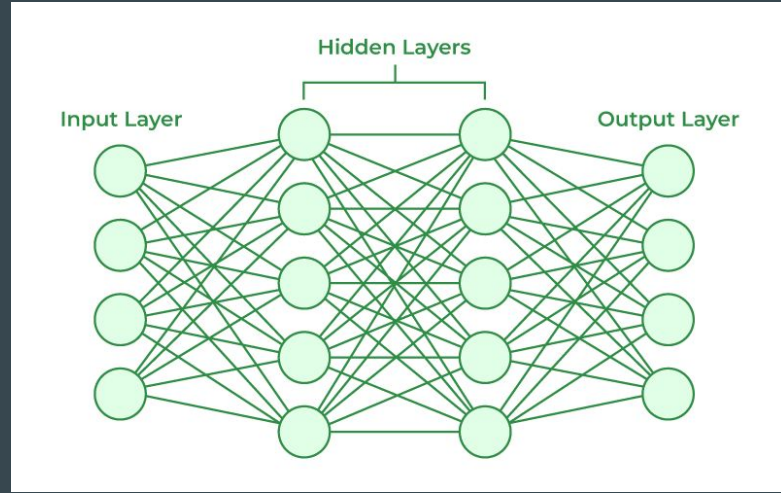
- 2 levels, with 128 nodes in level 1, 64 nodes in level 2
- TANH activation functions
- Trained over 1000 epochs.

Strengths

- Allows us to prediction/classification
- Can calculate classification rates.
- No basic assumptions besides assuming that the data are 'good'.

Weaknesses

- Adept at capturing interactions between features, but is not very interpretable.
- Technically possible to get variable importance, though difficult



Model Comparison/Selection - Classification Rate

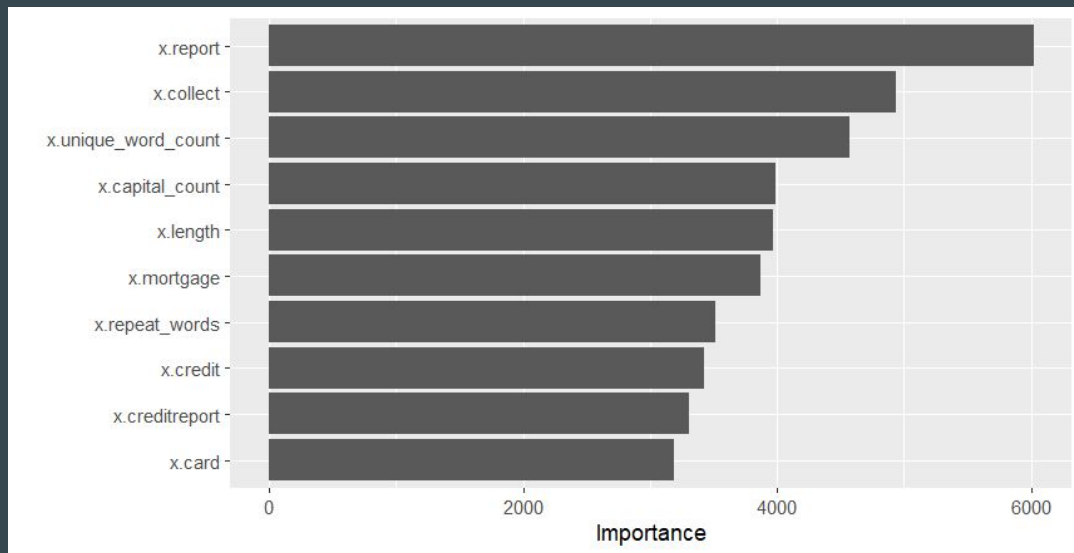
Model	In-Sample Classification Rate	Out-of-Sample Classification Rate
Model 1	78.9%	78.4%
Model 2	77.7%	74.8%

We will use Model 1 (Random Forest) due to:

- Improved classification rates
- Simpler explainability
- Ability to easily extract variable significance

Research Questions - Useful Variables

The word count, capital count and the length of the complaint message are useful for our predictions. Words in the complaint message such as 'report', 'reporting', 'collect', 'collection', 'mortgage', 'mortgages', 'credit', and similar words are also helpful in determining the appropriate department that a complaint should be forwarded to.



Research Questions - Commonly Confused Departments

- Payday, title, or personal loan complaints are often misclassified as mortgage or student loan complaints
 - This category is the most commonly misclassified
- Money transfer or money service complaints are often misclassified as checking or savings account complaints
 - The reverse is also true
- Debt collection complaints are frequently misclassified as credit reporting and consumer report complaints
 - Not as strong in the other direction

Research Questions - Unknown Complaint Classification

Using our chosen Random Forest model, we predicted the ten next complaints to belong to the following departments:

- [1] Credit reporting, credit repair services, or other personal consumer reports
- [2] Credit reporting, credit repair services, or other personal consumer reports
- [3] Credit reporting, credit repair services, or other personal consumer reports
- [4] Mortgage
- [5] Credit reporting, credit repair services, or other personal consumer reports
- [6] Credit reporting, credit repair services, or other personal consumer reports
- [7] Mortgage
- [8] Credit card or prepaid card
- [9] Debt collection
- [10] Student loan

Suggested Next Steps

- Better features
 - Looking online, there are some common features that could be included (with some significant work)
 - N-grams: ordered combinations of words
 - Official lemmatization - reduce words to their base form
 - Identifying parts of speech of keywords
 - Sentiment training (are parts of the complaint positive/negative, etc.)
- Consider industry, as certain industries may have different identifying features in what gets sent to each department

Appendix

...

Division of Labor

Caleb: Model 1, Results,

RJ: Model 2, Introduction

Both contributed to feature building, with Caleb doing a lot of the department profiles, and RJ doing the aggregation metrics (counts, ratios, etc.)