Assessment Report for AIA

KEE Consulting
Los Angeles

May 18, 2015

1 Data Infrastructure

Overall the data infrastructure is in good shape.

Greatest weakness of current dataset is that it is being populated by numerous agents each with varying standards. This poses two problems:

- The accuracy of the data becomes questionable:

  A bond with a "discharged" status may or may not have been forfeited. This makes the use of the bond status unreliable by itself.

- The variability of the data becomes unmanageable:

  For data analysis to work, non-numerical data must be categorized. This allows the mathematical models to translate each category to numeric form (i.e. Male/Female → 0/1), currently:

  - Charges associated with bonds are filled with either state-dependent crimal codes, abreviated crime, generalizations, or left blank altogether. A suggestion would be to have drop down menu with a severity rank of crime.
  - Defendant employement information is often filled with a business name or "self employed", which by itself is not very useful. A useful substitue could be an income bracket information from a drop down menu.
  - Defendant relationships are filled with strings such as "babby mama". Solution would be to provide options (i.e. ex-partner).

2 Project Roadmaps

2.1 Project A : A logistic regression model for Failure to Appear

2.1.1 The Goal

The goal is to construct a model which relates the probability of failure to appear (FTA) to variables through a coefficient for each variable. The Vision dataset is used jointly with the AIMS dataset.

A regression model: A statistical analysis used to predict scores on an outcome variable based on scores on one or more predictor variables.

Can be as simple as:

$$Y = B_0 + B_1X_1 + B_2X_2 + \ldots + \epsilon_1 + \epsilon_2 + \ldots \qquad (1)$$

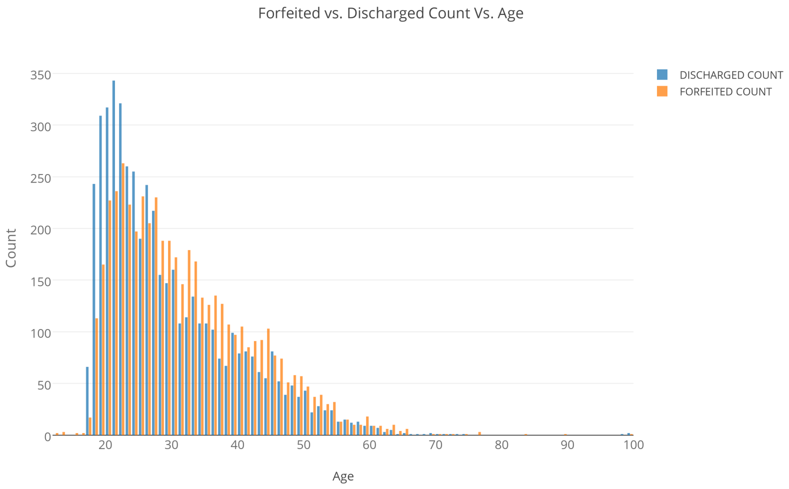
- Y: outcome variable (ex: Will fail to appear?)
- X: pridictor variables (ex: Defendents age, bail amount ...)
- B: coeffecients relating X's and Y
- ϵ: error terms (a.k.a residual)

2.1.2 Validity of model

As a proof of concept, four data variables were looked at for the initial model:

Characteristic of the defendant:

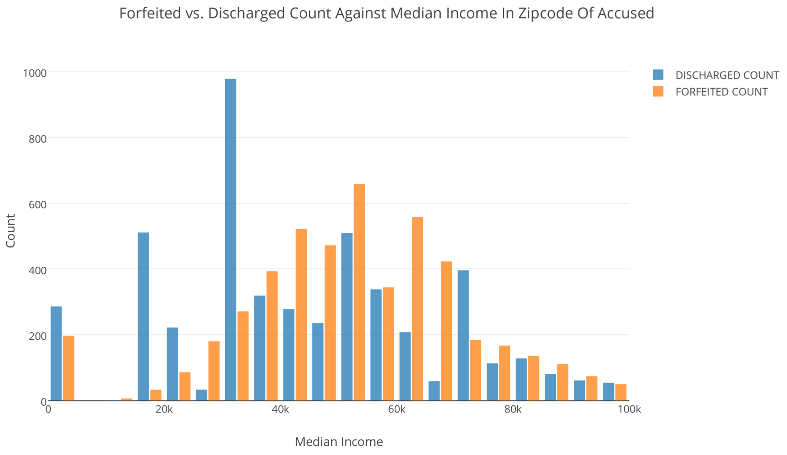
1. Age at time of the bond



2. Gender

Characteristic of the environment:

3. zipcode → income

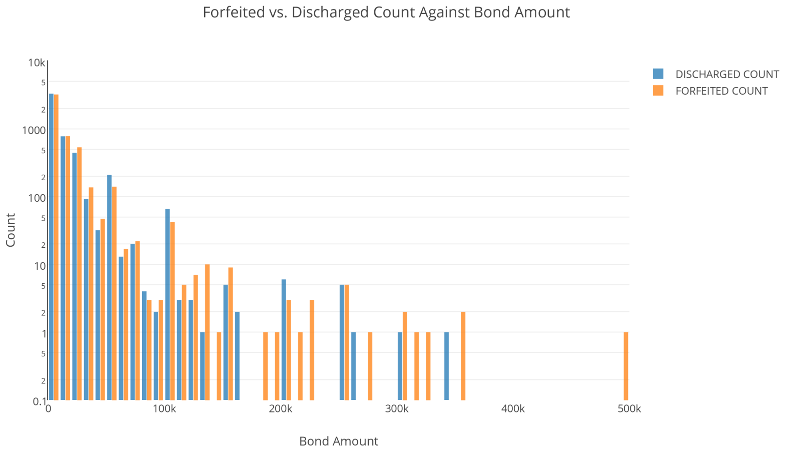


Forfeited vs. Discharged Count Against Median Income In Zipcode Of Accused

The average income for a zipcode was obtained through an api to the latest available U.S. Census.

Characteristic of the bond:

4. Bond Amount



Forfeited vs. Discharged Count Against Bond Amount

Running the Model:
Finding a relationship between X and Y which minimizes the model errors gives us:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3472  -1.0933  -0.7349   1.1296   1.8166

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.835339   0.099919 -18.368  < 2e-16 ***
catBond_Amount  0.013388   0.006470   2.069  0.03854 *
age             0.031466   0.002381  13.213  < 2e-16 ***
catZipIncome    0.183926   0.010761  17.092  < 2e-16 ***
genderM        -0.166691   0.053297  -3.128  0.00176 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

example model predicitions:

Defendent 1:

- Age: 38
- Gender: Female
- Bond Amount $35,000
- Zipcode Income $75,392
- Known to have failed to appeared

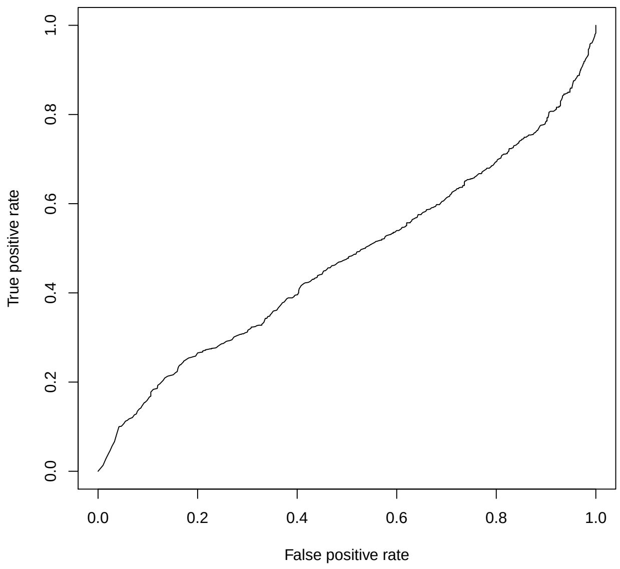
Probablity calculated by the model: 75% to fail to appear In reality, the bond was forfeited. This is called a "true positive".

Defendent 2:

- Age: 23
- Gender: Male
- Bond Amount: $5,000
- Zipcode Income: $101,905
- Known to have appeared

Probablity calculated by the model: 62% to fail to appear In reality, the defendant appeared in court and the bond was discharged. This is called a "false positive".

The aim is to maximize true positives and minimize false positives.



2.1.3 Deliverables

A “model” (equation) which would allow an agent to calculate the probablity of FTA given the characteristics of the defendant and the bond. The validity and performance of the model will be fully tested and reported.

2.1.4 Work Estimates

parsing/cleaning of entire training and test datasets categorizing variables (ex:work/collateral/co-signers) constructing model validation of model

1 month at $80/hour = $12,800

2.2 Project B: Agent Performance

2.2.1 Ranking and Reports

Ranking system: *Going beyond total premimum brought in:*

Goal: Construct an agent ranking system to assess the "health" of an agent's business. An agent with a low penal but good "health" could indicate a bottlneck in the business (i.e. premium rate, underwriting limit, BUF rate.) In order to build this ranking system an initial step would be to produce performance plots of agents and AIA. Performance plots could include three granularities, agent level, state level, and national level:

- premiums and BUF amount obtained from agents.
- Total penal written by agent
- granularity: agent, state, national
- comparison of these values by date ranges

2.2.2 Questions

Once the ranking metric is established, one can look for trends in the data as a solution to increase penal for an agent:

Example Question:

> *What contract variables contribute to premium collected? Is it Premium Rate? BUF Rate? Underwriting limit? Others?*

Testing "Premium Rate":

The following figures show the premium collected from agents throughout their numerous contracts with AIA:

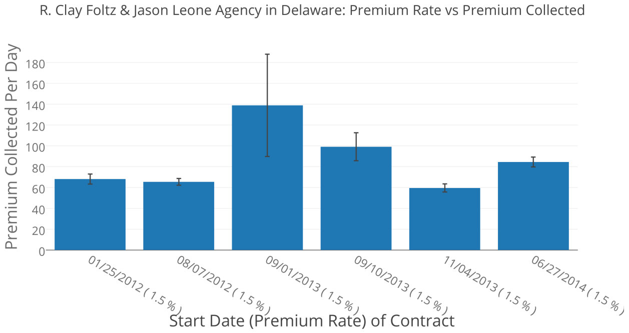
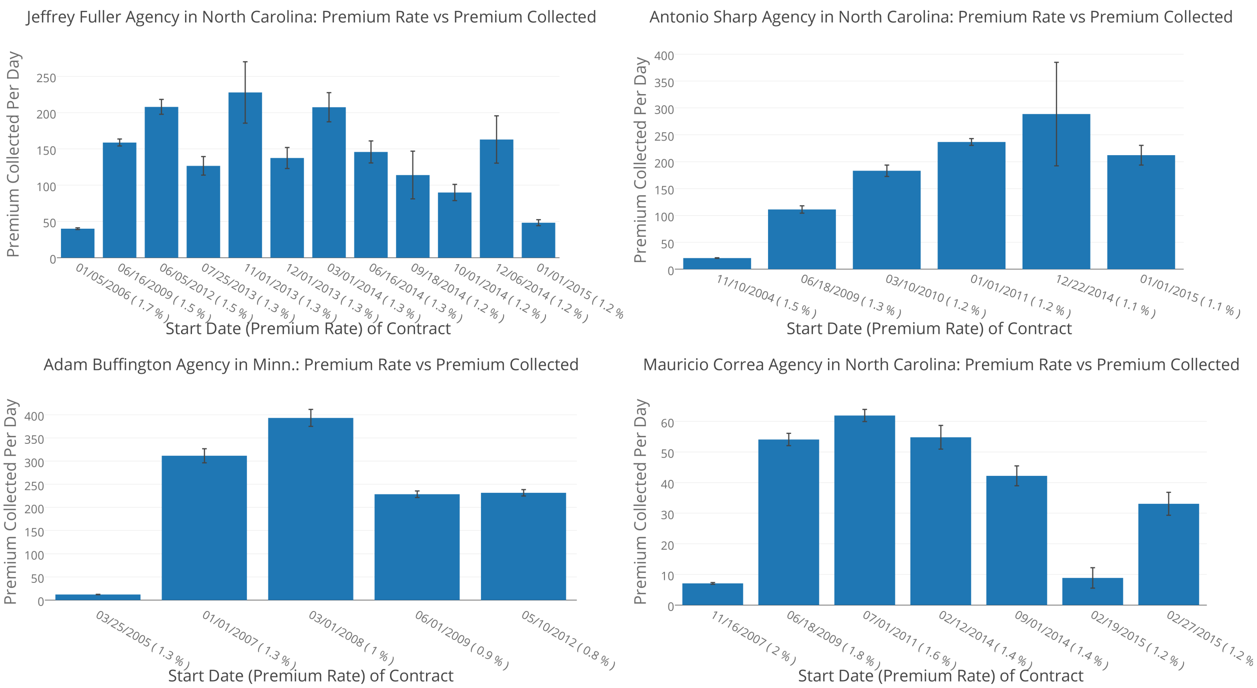


Figure 1: An agent with a constant premium rate



Jeffrey Fuller Agency in North Carolina: Premium Rate vs Premium Collected

Antonio Sharp Agency in North Carolina: Premium Rate vs Premium Collected

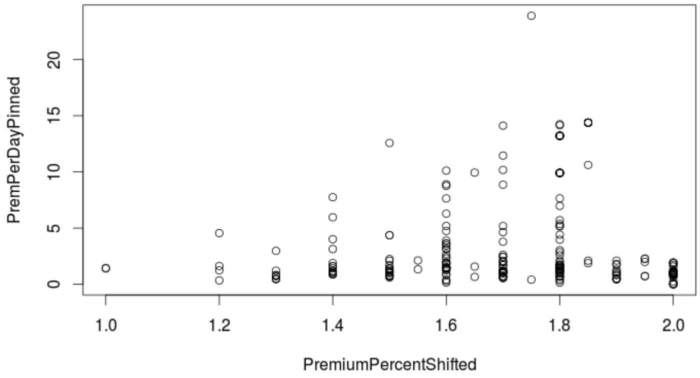
Adam Buffington Agency in Minn.: Premium Rate vs Premium Collected

Mauricio Correa Agency in North Carolina: Premium Rate vs Premium Collected

Figure 2: Agents with varying premium rates

Looking at the different agencies, it becomes apparent that on average each write varying volume of penal per day.

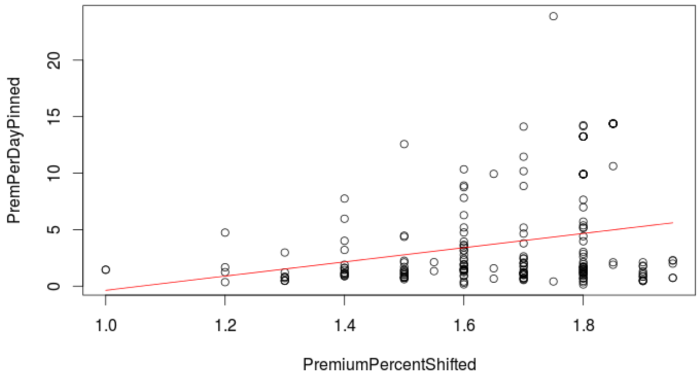
We can construct a linear model which relates the premium collected from agents as a function of premium rate. Plotting these two values against each other we get:

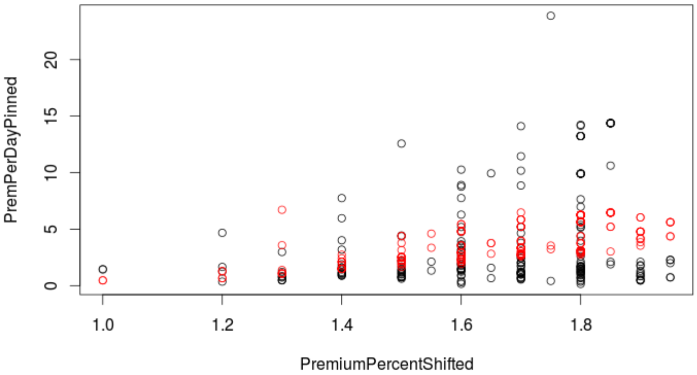


```
Residuals:
   Min     1Q Median     3Q    Max
-4.875 -2.978 -1.478  1.250 19.538

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             -6.677      2.694  -2.478  0.01406 *
PremiumPercentShifted    6.299      1.613   3.905  0.00013 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.266 on 197 degrees of freedom
Multiple R-squared:  0.07183,Adjusted R-squared:  0.06712
F-statistic: 15.25 on 1 and 197 DF,  p-value: 0.0001296
```

Adding another variable...

```
Residuals:
   Min     1Q Median     3Q    Max
-5.875 -2.808 -1.090  1.548 20.339

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           -4.885e+00  2.634e+00  -1.854   0.0652 .
UnderWritingLimit      1.253e-05  3.118e-06   4.020  8.3e-05 ***
PremiumPercentShifted  4.102e+00  1.648e+00   2.489   0.0136 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.111 on 196 degrees of freedom
Multiple R-squared:  0.1425,Adjusted R-squared:  0.1338
F-statistic: 16.29 on 2 and 196 DF,  p-value: 2.856e-07
```





In some fields, it is entirely expected that your R-squared values will be low. For example, any field that attempts to predict human behavior, such as psychology, typically has R-squared values lower than 50%. Humans are simply harder to predict than, say, physical processes.