Machine Learning

Andrew Smith, Kofi Otseidu, Manav Trivedi

November 6, 2018

1 Predicting which officers will recieve complaints

For this we set out to predict which active officers would receive a complaint in 2017. There were some problems with this that I'll get into later. First we created a table that included each officer's race, gender, rank, and number of complaints prior to 2017. From there we used a decision tree model to attempt to classify the number of complaints an officer would receive in 2017 (zero, one, or two). Our model preforms with an accuracy of 97% which seems good. Ultimately there are around 10,000 active officers in Chicago. Approximately 100 of those received a complaint in 2017. The model effectively states that any given officer has a 4% chance of getting a complaint in a given year, and that varies slightly (± 0.1%) with age, gender, rank, etc. The defining feature is the prior complaints, which makes sense, as for example, an officer with 12 complaints now has a probability of 20% of receiving a complaint in 2017.

Going forward, while the percentage is interesting to note, there are definitely better ways to set this question up. We'd most likely be better off evaulating whether or not an officer will receive a complaint over his or her career as opposed to for a year. For that it would be helpful to include the district/neighborhood, salary, and awards information, although they would be tricky to include.

2 Predicting the cost of a settlement

So for this problem, we first create a feature matrix. Included as features are the officer rank, age (at time of settlement), gender, race, awards, neighborhood, and number of complaints and settlements prior to the current settlement. From that we use a regression to try to predict the cost of a settlement. We used a generalized linear model with a log-gaussian fit. We used cross validation to refine the hyperparameters and output the best model.

Looking at the result on test data, the predicted data distribution generally matches the actual distribution. Below are some of the plots, and the code will be up on our github.

The results, as mentioned, generally match the actual settlement values. However it would probably be worth continuing to tinker with the model and other regression models to determine whether we have the best fit here. We'd also like to note that for the purposes of this analysis settlements with multiple officers are effectively counted more than once, as each officer will show up in the data. To fix this in the future we could scale the settlements (i.e. a \$10 million settlement with two officers would be split into \$5 million settlements), but that has it's own problems. We could also just combine the stats of the officers into one row, but that again would introduce some issues, so for now at least we left it as is.

Looking at the error plot for the output test values, the model does well for smaller valued settlements but has problems with the larger settlements. This isn't unexpected, the larger values

Figure 1: Predicted Settlement Values

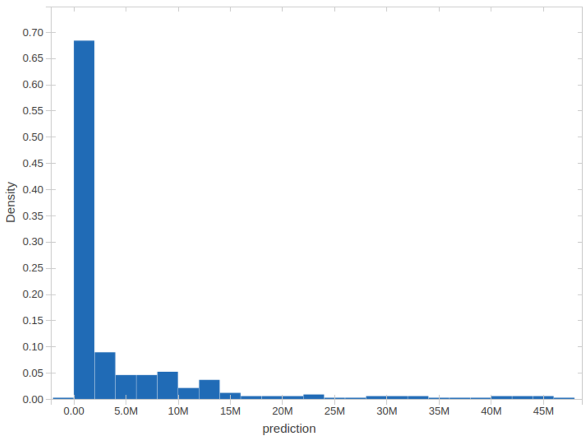
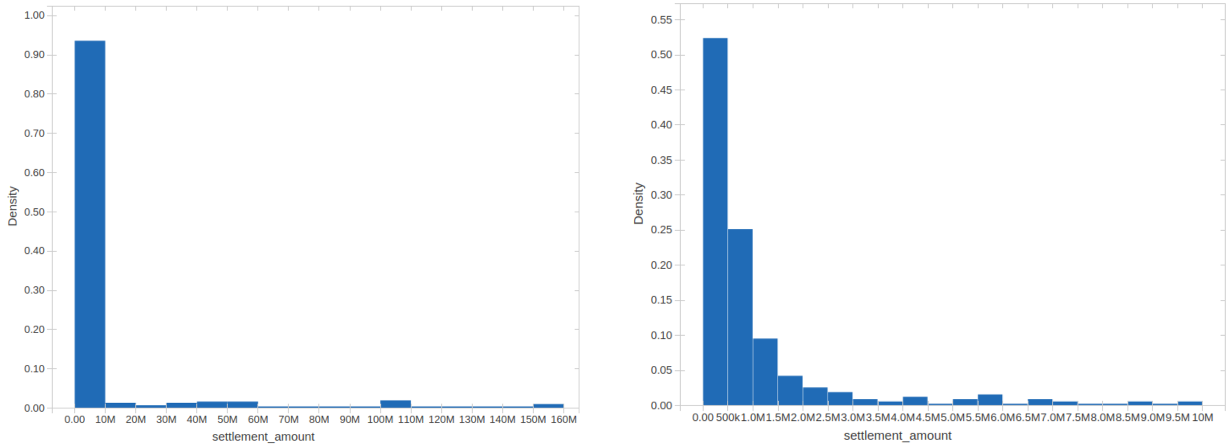
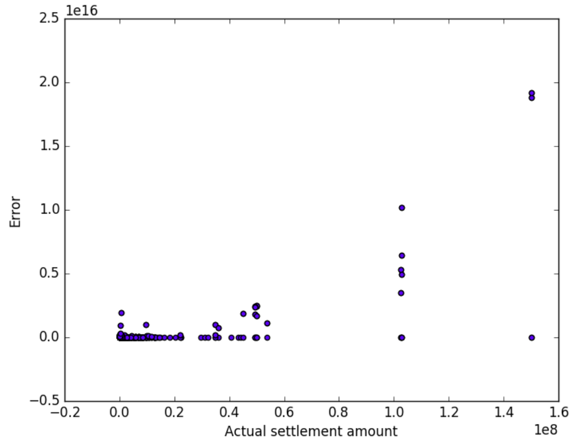


Figure 2: Actual Settlement Values



might be considered outliers and so don't fit in the model, but there are enough that I would be wary to exclude them.

Figure 3: Error on predicted values


1e16
Error
Actual settlement amount
1e8