Checkpoint 4: Machine Learning

The Creative Wolves

Masum Patel, Sizheng Zhang, Jiawei Zhang

Introduction

During the first three checkpoints, we dove deep into the attributes of officers who received the most complaints in the force and categorized the officers into two groups. The top 4000 officers who have received the most complaints are called the "Repeaters" and the officers who have received at least one complaint are called the "Offenders". Our theme was to explore and identify commonalities among the "Repeaters" and the first two checkpoints have shown that higher salary and awards are received by the "Repeaters" than those received by "Offenders" on average. Thus, in the fourth checkpoint, we want to build machine learning models to predict the number of allegations received by repeaters to see whether high salary and awards really indicate a large number of allegations.

Salary, Awards - Number of Allegations

Model Results

Regression

We decided to use a regression model to predict the number of allegations that repeaters might have based on different criteria. We have two regression models for comparison: linear regression and gradient boosted tree regression. We applied both of these two models to two different datasets: one contains the salary and number of awards received by repeaters and one that does not contain such information, to

investigate the effect of these two properties on our prediction. The results are shown in the figures below.

```
[Running] python -u "/Users/owenzhang/Desktop/Northwestern/2020-2021/Fall/rogers/cp4/Models.py"
[0.859331   0.86105807 0.8745039  0.86092071 0.85516395]
Linear Regression (with salary and num_awards) average 5-fold CV R2 score: 0.8621955275625961
[0.97258366 0.96714124 0.96914507 0.95988792 0.96639117]
GBT Regression (with salary and num_awards)  average 5-fold CV R2 score: 0.9670298110526575
Best parameters are: {'learning_rate': 0.2, 'max_depth': 3, 'n_estimators': 200}
Best score is: 0.9794698311654683
```

Figure 1: The performance of two models on the dataset with salary and num_awards

```
[Running] python -u "/Users/owenzhang/Desktop/Northwestern/2020-2021/Fall/rogers/cp4/Native_models.py"
[0.85333981 0.85520182 0.86601543 0.86193868 0.84703287]
Linear Regression (without salary and num_awards)  average 5-fold CV R2 score: 0.8567057205186845
[0.96733058 0.95717142 0.96399424 0.95175834 0.95681773]
GBT Regression (without salary and num_awards) average 5-fold CV R2 score: 0.9594144615615822
Best parameters are: {'learning_rate': 0.2, 'max_depth': 3, 'n_estimators': 200}
Best score is: 0.9731239284141611
```

Figure 2: The performance of two models on the dataset without salary and num_awards

Co-complaints/year, Awards, units - Number of Complaints

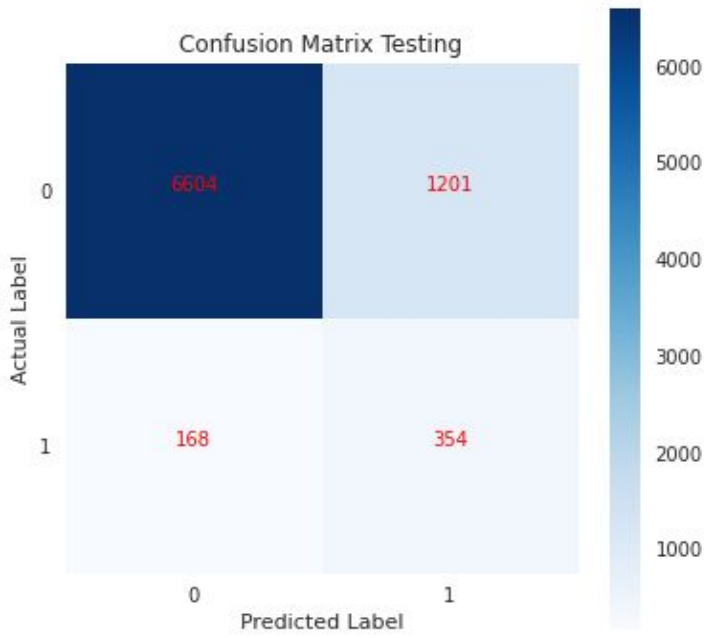
Model Results

Classification

We decided to use a classification model to predict if the repeaters will have a complaint next year based on information from this year. What we do is, we grab data of pairs of officers among repeaters and see how many co-complaints they have, number of total awards they got and whether they were in the same unit in a particular year. And we try to predict if that pair is likely to get a complaint together next year.

We then use some feature engineering and grid search and cross validation using catboost and predict results.The results we get for classification are as follows:

```
Train Accuracy : 0.837
Validation Accuracy : 0.822
Test Accuracy : 0.836

Train Classification Report                    :
              precision    recall  f1-score   support

           0       0.97      0.85      0.91     24680
           1       0.26      0.68      0.38      1964

    accuracy                           0.84     26644
   macro avg       0.62      0.77      0.64     26644
weighted avg       0.92      0.84      0.87     26644


Validation Classification Report                    :
              precision    recall  f1-score   support

           0       0.96      0.84      0.90      6150
           1       0.24      0.62      0.35       511

    accuracy                           0.82      6661
   macro avg       0.60      0.73      0.62      6661
weighted avg       0.91      0.82      0.86      6661


Test Classification Report                    :
              precision    recall  f1-score   support

           0       0.98      0.85      0.91      7805
           1       0.23      0.68      0.34       522

    accuracy                           0.84      8327
   macro avg       0.60      0.76      0.62      8327
weighted avg       0.93      0.84      0.87      8327
```



Analysis

- From the results of the regression models, we can tell that taking salary and the number of awards received by repeaters does not influence the accuracy of our prediction by a lot. Before adding these two criteria, the R^2 score of our linear regression model is 0.856, and that of our GBT regressor is 0.959. After adding salary and num_awards into our predictors, the R^2 score of the linear regression model increased by 0.08, and that of the GBT regressor increased by 0.075, which is not a lot of variation. Thus, we can conclude that even though salary and number of awards seem to be significant predictors of the number of allegations received by repeaters, other predictors like age, civilian allegation

count, and complaint percentile seem to be sufficient in predicting the number of allegations that an officer has.

- From results of our classification models, we see that our model gives high recall and low precision values, which basically means it gives results but most predictions it makes are incorrect. For our example, if the model predicts a pair getting in trouble next year, there's a 68% chance it will get into trouble. At the same time 15% of times, when it predicts a pair getting into trouble next year, it may not. So we do see, intuitively and with the numbers that there may be some correlation between co-working patterns and co-complaints.

Conclusion

- To conclude we see that we don't find any strong patterns between officers' co-working patterns and co-complaints. However it may also be because of some of the issues like:
    - We took only the repeaters to look at that hypothesis. We could further see by checking more number of officers.
    - Also taking into account other features like race, gender, seniority, age, etc along with the co-complaints and where they work may provide us some insight on the correlation
    - We carried out the analysis and predictions looking at a pair of officers. It may be possible that as the number of officers increases, we may get a stronger correlation between complaints and co-working patterns as we have already kind of seen with the Crews dataset.
    - We also had to tackle sme data problems like missing values of units the officers were assigned in some years. Maybe more data could have led us to better insights and understandings.