# THE OAKLAND A'S FINAL WRITE UP

***Prepared by:***
*Marie-Reine Axalan,*
*Marcos Fontes*
*Jordyn Gerstle-Goodman*

*December 8th, 2021*

# DELIVERABLE OUTLINE

Within this document, executives can understand where our team has identified a problem and why we have developed a model for the data we extracted to better understand where we can level and justify the judicial system.

01

## BUSINESS UNDERSTANDING

- Business Problem
- Value
- Business Evaluation

02

## DATA UNDERSTANDING

- Data selection and understanding the data we found

03

## DATA PREPARATION

- Specifications regarding how our data was integrated to the analysis

04

## MODELING

- Understanding how we developed our model
- Analytics tool that we dictated where best for the data analysis
- Translating how our model will solve the business problem

05

## EVALUATION

- Model and analysis evaluation
  - Business Case
  - Potential and implementation

06

## DEPLOYMENT

- Putting results into action
  - Considerations and risks
- Action plan/roadmap

# 01 Business Understanding

The criminal justice system in America is notoriously known as an unjustified system that enrages unjust outcomes for all stakeholders involved including the entire public. In a liberal democracy, the criminal justice system employs state-sanctioned violence to discourage and punish those who are guilty of crimes against individuals and the state. Over-criminalization would have never become the problem it is today if all criminal charges were unbiased using extensive and costly mechanisms of trial. A system that features countless constraints against minorities that are exploited by criminal charges that directly incite them to unnecessary sentences.

With an issue that enormously reaches across an entire justice system, our team has had the opportunity to begin taking a step towards the right direction of correcting an unjust system. Taking our own initiative to approach the problem where it starts in the trial process. Our team has developed a strategy that will allow us to control how citizens are fairly criminalized based on their actions and not on any other variables. While only serving as a piece of the ever-going problem that is the justice system, our group has developed a solution to an influencing variable of the ever so large  imprisonment issue America faces today.

## THE PROBLEM...

### BAIL REFORM AS A BUSINESS

After an individual is arrested - rightfully or wrongfully - someone's ability to leave criminal detention and return to their homes & oppose the conviction is completely dependent on their ability to pay bail. Bail, simply put, is the amount of money that defendants pay based on a posted amount where they would be released from detention of a crime until their trial. Something that doesn't serve as a fine or penalty will allow defendants the opportunity to appear for trial after being arrested... But for those who are unable to pay the posted bail on their conviction it is not as privileging as it sounds.

# THE PROBLEM *Continued*

When someone is unable to pay a posted bail on their criminal convictions, the defendant is immediately perpetuated into an endless cycle of jail time and jail-indicted poverty. With 1.6 Million people incarcerated, there are an additional 600,000 individuals who are jailed in local jails throughout the US. With 70% (*PrisonPolocy.org*) of individuals who are in jail being held on pre-trial who are legally presumed innocent, there is obviously an issue. While bail was developed as an equalizer freeing defendants until their trial date, it has ultimately developed another bias in the judicial system against the poor (one could argue specifically marginalized minorities). A development that was originally determined on:

- the risk of the defendant fleeing
- the type of crime alleged
- the "dangerousness" of defendants
- the safety of the community

more people find themselves in jail because they simply cannot afford the price of bail. While bail does serve as a tactic to help defendants get out of jail, the concept has transformed into something that is more of ruthless business practice. As an entity of the public, our team has decided to focus on the underwriting that comes with pre-trial bail and how it can further cleanse the judicial system and imprisonment as a whole.

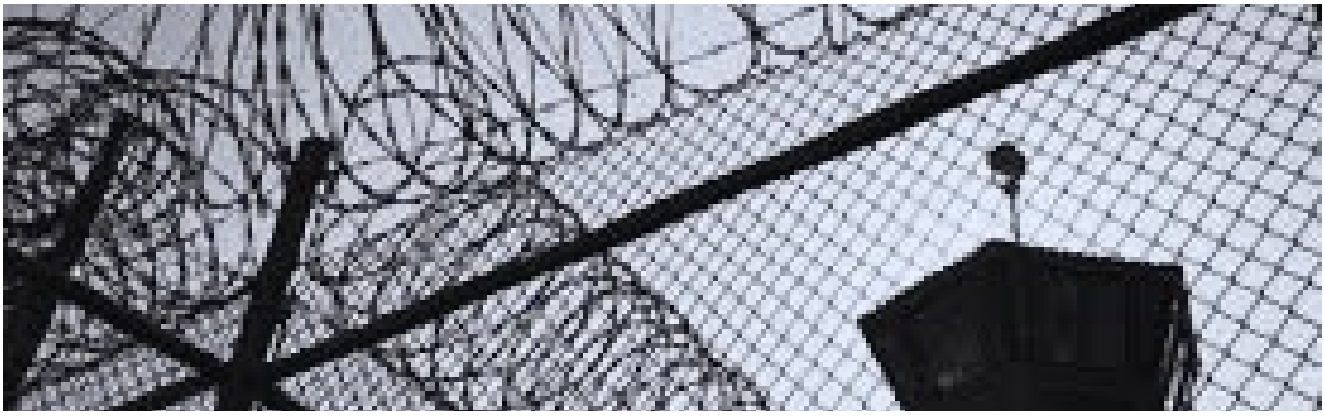
According to Kate T. May 'African-American and Hispanic people are more likely to be arrested, more likely to be issued bail, and less likely to be able to afford it.' (*Ted.Com*). Following this, every day someone is in jail, the more likely they are to lose their job, home, citizenship status, and even custody of their children fueling a pipeline of criminalizing poverty and in the moment direct conviction of a crime. The backend of the judicial system not only pushes unrealistic bail amounts but more importantly drives defendants to be guilty even if they aren't. While bail serves a greater purpose of freedom in a country that was built off the term it might be time to rethink the idea.

# VALIDATING THE ISSUE...

## THE VALUE TO THE PROBLEM

This bail issue is directly linked to the bigger problem of prison overcrowding and criminalizing poverty and our team's solution to an unjust entrance of the system might just be a step in the right direction. Considering the existing biases that impact bail amounts, the individuals who are facing the judicial process and the incarceration system itself are the ones who are the most impacted and targeted. The results in this analysis will be able to provide our team, audience and executive stakeholders with insights as to how bail amounts correspond with existing biases that influence who exactly is being incarcerated and why.

Assuming that demographics play a large role in the amount of bail presented in each trial, we are hoping to identify the inequality and prejudice that is currently influencing the incarceration of almost two million people. While ultimately external subsequent policy makers will play a leading role in how criminal justice reform is influenced by this analysis, we believe that our discoveries will play an important role beyond posted bail. In the path going from arrest to trial detention, it's presumable that there are a number of ways that individuals can influence the final outcome, but for 50% of adults who can't even afford a $400 emergency expense the truth holds otherwise. In the bigger picture the directions of our analysis will portray clarifying insights on the confinement system itself, overall assuring all convictions are equal and flatten the prison overcrowding and criminal poverty issue fronting our country.



# EVALUATING THE PROBLEM..

## GOING IN THE RIGHT DIRECTION

By investing in this statistical investigation of bail reform, we are hoping to decrease the basis present in incarceration by highlighting inequalities and setting fairer terms assigned to each individual defendant.

With our Analysis we will be able to target the correlation between demographic information and the offenses committed with the amount of bail posted, eliminating any sort of bias that is against minorities and marginalized communities. Reforming the bail system will reap many benefits that can be evaluated by measuring such as less overcrowding in prisons, lower poverty rates among marginalized groups and generally preventing innocent people from spending their lives in the prison system. As a result of bail reform comes the decrease in taxes for taxpayers, given the ability to decrease the amount of detainees. Lastly, by allowing people a true chance at justice without the heavy burden of thousands in bail amounts we can limit the amount of people in prison. Lowering the costs prisons need to pay per inmate which would decrease the amount of tax money that needs to go into prisons.

All of these improvements can be evaluated with the implementation in the pre-trial posted bail underwriting and will ultimately let us know the effectiveness of our analysis. Following through with any solutions developed after our analysis we can assure that with further research and legislation, we can redevelop a justice system that is not only equitable but just.

# 02 Data Understanding

‘Overcrowding is a consequence of criminal justice policy not of rising crime rates..’ as stated by the Penal Reform International non-profit. As our group has developed our own strategy to understand how pre-trial bail influences the surplus of imprisonment that undermines the ability of prison systems to meet basic human rights such as healthcare, food, and accommodation. We found a general overlap with the groups initiative also motivated by the excessive use of pre-trial detention, in which we were able to find and extract a data of source that would allow us to further investigate where and why the pre-trial bail system is currently flawed, inciting systematic bias and lifetime criminalizing poverty. After finding our way to the state of Connecticut's Open Data profile, our team was able to find a dataset recording individuals who at the time where being held in Department of Correction facilities while awaiting trial. Deciding that this would be an appropriate sample to start our analysis we would use this sample for our analysis. Provided below is an outline of the data we found and used for our analysis:

| DATA TYPE | FIELD NAME | FIELD DESCRIPTION |
|---|---|---|
| INT | IDENTIFIER | Individual Inmate Identifier |
| CHR | LATEST ADMISSION DATE | Most recent date in which the inmate has been admitted |
| CHR | RACE | Race of Inmate |
| INT | AGE | Age of Inmate |
| NUM | BOND AMOUNT | Amount of bond for which the inmate is being held. |
| CHR | OFFENSE | Offense for which the bond amount has been set. |
| CHR | FACILITY | Correction facility Department |
| CHR | DETAINER | Enforcement person who detained Inmate |

Whilst we found ourselves working with an intimidatingly large data set, we found immediate opportunities as to how we could develop an analysis that would allow us to demonstrate the business problem we selected and how it could be accordingly engineered. With this Dataset our group has decided to Analyze the correlation between race and offense with the bond amount. We hope to identify a correlation as well as find an algorithm that would predict the bond amount given the severity of crimes. As we have proceeded with our project we have concluded that our analysis will be labeled as a supervised learning model focused on the Bond amount and seeing if demographic features, i.e. Race, Gender, Age, etc. factor into that overall amount. Expecting to see if any correlation of these demographics would give us an idea as to whether prison overcrowding can be assessed at the root of the problem that is pretrial detention .

## 03 Data Preparation

Our original data set began with over 5 million rows, which wasn't ideal for our project in terms of efficiency. Though our dataset records crimes committed in New England since 1980 with the type of crime committed, facility the person charged was detained in, the admission date and other demographic information about each detainee. Though having a larger dataset indicates better performance of our model, our machines were not powerful enough to run all this data multiple times. Due to this reason we resolved to remove data instances that were before 2018, accounting for the relevance of our model.

We chose the bond amount as our predictor variable and after visualizing this data we noticed that this feature was very right skewed. To transform the data, we square rooted the bond amounts to approach a normalized curve. In our data preparation, we bucketed the bond amounts with 5 different classes ranging from "Very High" to "Very Low". We chose to do this so we could run classifiers with a discrete variable since Bond Amount is a continuous variable. After consulting with Professor Zhang, we realized that this was not the best practice as it would be removing valuable data points and our model would not be as precise. By using the continuous bond amount variable, we were limited to regression models for our project.

In our first round of data preparation we split 75% of our data set to the training set and the rest to our test set. This left about 2 million instances in our training and 700k instances in the test set. When we started running models on this training set, the file was too big which prevented us from any models on our machines. To resolve this problem we randomly subsetted our training and test data by 50%, which allowed us to run a linear regression and KNN models on this dataset.

Regarding the actual features of the dataset, several points of data included multi-class classifiers. For example, types of offenses had over 200 unique values and would prove difficult to interpret its effect on the predictor variable. For this value we sorted each offense between felony, misdemeanor, and other. If we had a better understanding of the data and the nature of crimes, we believe it would have been useful to include if a crime was violent or non-violent.

# 04 Modeling

We focused on building linear regression models and K-nearest neighbors. The linear regression models contain different predictors; one group of the models will attempt to predict the specific bond amount (a continuous data type), while the other group will attempt to predict a bucketed bond amount. The bond amounts that are bucketed are classified as ‘’Very Low’, ‘Low’, ‘Moderate’, ‘High’, ‘Very High’ to apply a label to the type of individual’s bond amount.

The linear regression model predicts the bond amount based on the features of our set. As explained in our data preparation section, by keeping our bond amount feature continuous, we were limited to regression models. This allowed us to keep a majority of our data points rather than bucketing them into categories. We mainly used our regression model to analyze our features and their interaction with the predictor variable. With our first run of the regression model, we included the offense type which had 273 different values. Including this made the model hard to interpret, thus we decided to sort the offenses by class (felony, misdemeanor, and other).

# MODELING *Continued*

After transforming the data in this way, our linear regression was much easier to interpret. Interaction terms also proved to be important in creating the regression model. An interaction between race and offense was discovered, indicating that the offense type committed affected the bond amount differently when race changes. Though we are aware that other variables that we are not aware of could cause this change, we believe this to be an important factor as this interaction was statistically significant in the model. We also used linear regression to analyze the significance of each feature. Statistical significance was found in a majority of the features in our model, which proved to be beneficial in our analysis.

The KNN model assists in using the data points to separate into different classes. This non-parametric algorithm works for a project like this based on the fact that it will assist in informing our exploration of bias in the setting of bond amounts by grouping the data points into different segments. Given that our data type of the predicted value is a continuous data type, we implemented a regression based K-Nearest Neighbors. The model structure will be completely informed by the data and slightly opposes the ideology of a linear regression that adheres to theoretical assumptions.

In order to accurately use our KNN model, the columns with labels needed to be transitioned to their own features and used as a binary predictor. For example, taking the RACE column that listed white, black, Hispanic, etc. and transforming that into a WHITE column, BLACK column, HISPANIC column, etc. where they will be marked with a 1 if positive class. The model is trained with the following columns as X: 'WHITE', 'BLACK', 'HISPANIC', 'ASIAN', 'AMER IND', 'MALE', 'FEMALE', 'AGE', 'MISDEMEANOR', 'FELONY', 'OTHER OFF'. The target is Bond amount which aligns with the aforementioned linear regression models.

# ANALYTICS TOOLS

For our project we used Rstudio and Python to create and deploy our models. Rstudio and Python were the best tools for analytics because these are the coding environments we felt the most comfortable in. Rstudio was a useful tool because it included the needed package and easy functionality to perform our linear regression. We found it easy to analyze features and adjust our model accordingly. Through this program, we have learned to analyze interaction terms through R, which helped us better understand the variables. Plotting features was also very helpful to our analysis and we were able to easily do so in Rstudio.

The biggest con of using Rstudio was that it was much slower with our large dataset. While preparing and running our model, we had to restart R multiple times as it would break down easily. Along with the program breaking down frequently, Rstudio was much slower than other coding environments. We would have to wait for as long as an hour to let Rstudio run our code, which interfered with time we had available. Another con included the various packages we had to install to make everything work. This required a lot of time researching and referencing the R manual to find what we were looking for. For our KNN model, we felt more comfortable performing this model in python. Python is frequently the most performed coding language for machine learning. Using this for our KNN model proved to be easier to run and understand compared to Rstudio. A few cons with using Python specifically in Jupyter notebooks included the long runtimes and the various packages. This is very similar to what we found with Rstudio and could largely be attributed to our large dataset.

Other analytics tools we could have used would be Alteryx which we believe would have been a better alternative for this model. Due to the restriction of this project being mostly in R, we did not have the time to explore the possibilities of this tool. Though Alteryx has very many machine learning capabilities that our team is very comfortable with using, we would imagine that the run time would have been even longer than what we experienced with Rstudio. Data Robot would also be a great tool for this analysis, and something we mentioned in our original model. Data Robot is a great tool to find the best model for our dataset, and performs faster with its predictive modeling than other tools.

# 05 Evaluation

The results of our data analytics couldn't be evaluated in a business sense using tools like ROI. The point of our analysis would add more income to the bail system but rather ensure that the people in the system are treated equitably. Reforming the bail system reaps many benefits that can be evaluated such as less overcrowding in prisons, lower poverty rates among marginalized groups or preventing innocent people spending their lives in the prison system. Another improvement as a result of bail reform comes from the decrease in taxes for taxpayers if we are able to decrease the amount of detainees. By allowing people a true chance at justice without the heavy burden of million dollars bail amounts we can limit the amount of people in prison. This would lower the costs prisons need to pay per inmate which would lower the amount of tax money that needs to go into prisons. All of these improvements can be evaluated with the implementation of our model and let us know the effectiveness of our analysis.

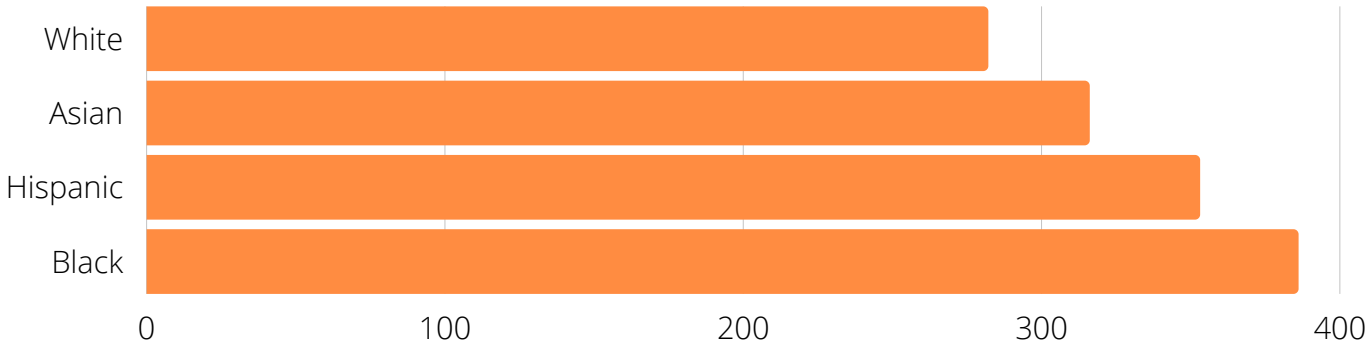
## LINEAR REGRESSION MODEL

Our linear regression model provided a baseline for analyzing the features of our dataset. We were able to build a model using detainer, offense type, gender, age, and race as our features. This model produced an RMSE of 166.35 on our test data, which for accuracy, we would be comfortable implementing. Significant variables included race, specifically black and white instances. The significance of this variable was further analyzed in the interaction term explained below.

The offense type will have a varying effect on the bond amount when race is different. After observing the interaction terms, there is a clear interaction between race and offense. Our findings show that there is an interaction between race and offense. When isolating race and offense in our model, a subject being black and committing a felony has a bigger effect on the final bond amount than other races (See Appendix A Fig. 1) Thus, we believe an interaction between these two variables was necessary in our model.

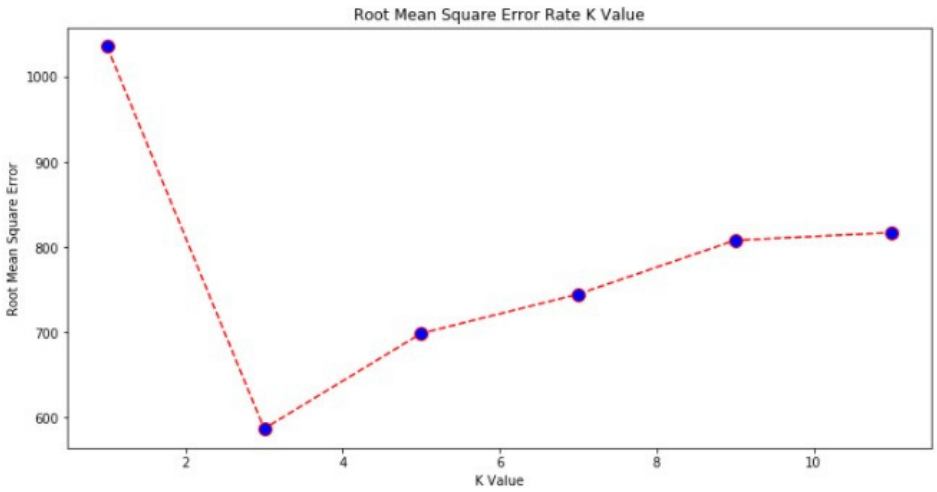
Looking at the coefficients in the model with the interaction term included (See Appendix, Fig. 2), we found that being black and being white had opposite effects on the bond amount and where significant variables in our model. We believe this is an interesting insight to consider when evaluating the biases of our model. We will discuss in detail the risks of having significant demographic variables in our model. Another significant variable to consider is the impact gender has on the bond amount. Being a male had a significantly large impact on the bond amount when considering female subjects. This could prove as a point of interest in future analyses. Lastly, the type of offense was another significant feature and a feature we believe is important to include in our analysis. As discussed prior, bail amounts should be defined by the crime itself and the criminal history rather than demographic information regarding the person committing the crime. This features significance was reasonable in the lense of our predictor variable.

Below an initial analysis was run looking at the median bond amount for each race for felonies, which lead us to explore the interactions this feature has with offense type.



## KNN MODEL

The purpose of initially implementing the K Nearest Neighbors model was to look at feature similarity and eventually be able to predict new values of data given the features we currently had. When implementing the KNN model, it was found that the lowest Root Mean Square Error it could obtain was 586.84 (Figure 3) with a k of 3, but given that the range of instances for the target feature was from 1 to 758, this model was not as predictive as we would hope. We intended to use the demographic features of the detainees, i,e. ethnicity, type of offense, and gender. The R2 score was negative when using these features showing that the demographic features alone did not actually correlate to the bond pricing as much as we hypothesized.




# 06 Deployment

## DEPLOYING THE RESULTS

The main point of our analysis would be to bring attention to the biases that affect the bond.amount. The future goal of this project would be to refine our model and remove the inherent racial biases that we discovered. The deployment would be for this model to serve as a starting point for significant change in the bail system.

# DISCLOUSURES AND ADVISING

The entity that would deploy our recommendations should be aware of the biases model and the possible error. Other variables may be useful to include in the model if our team were able to find them and include them in our analysis. The possible variables could be whether the crime committed is considered violent or non-violent and whether this is the first time the person was detained. Using more features that focus on the crime(s) committed, rather than the demographic information of the person committing the crime would call for a less biased model.

# CONSIDERATIONS

Given that the situation these are imperative to are legal proceedings, 'do no harm' is one of the largest ethical considerations. The current evaluation process for the bonds causes unintended harm to the participants, and even with the implementation of this model, it may still continue to do harm as the legal system itself is built upon institutions that promote the intended harm of sub groups. The ethical implications come into play specifically with our model because it uses demographic information to assess the problem. Given the potential of a bias of race in setting bond price amounts, if we are trying to mitigate this issue, a potential bias may be created while trying to avoid the initial one. If results are saying that there is an issue with a specific race being a significant factor in setting the amount, then trying to oppose that specific feature could sway the distribution towards other factors or ethnic backgrounds.

# MITIGATING THE RISK OF THE FUTURE

Features associated with the crime would be better indicators as opposed to demographic features. When using race or even items like gender as a feature could lead to problematic assumptions and biases in models. One of the biggest risks is that it would create legislative strife and issues on a social level. With enough evidence to present the initial findings of these models and gathering support on the implementation of the matter could assist in support and mitigation of any disapproval. We would want to be able to implement a wider scale model over a more extended period of time with more features in regards to the actual crime itself, not the personal characteristics of the individual. Our biggest risk is the level of acceptance with a data set that is specific to the New England region.

# Bibliography

Initiative, P. P. (n.d.). Detaining the poor: How money bail perpetuates an endless cycle of poverty and Jail Time. Detaining the Poor: How money bail perpetuates an endless cycle of poverty and jail time | Prison Policy Initiative. Retrieved December 7, 2021, from https://www.prisonpolicy.org/reports/incomejails.html.

May, K. T. (2018, August 31). How the bail system in the US became such a mess - and how it can be fixed. ideas.ted.com. Retrieved December 7, 2021, from https://ideas.ted.com/how-the-bail-system-in-the-us-became-such-a-mess-and-how-it-can-be-fixed/.

Prison overcrowding. Penal Reform International. (2021, May 20). Retrieved December 7, 2021, from https://www.penalreform.org/issues/prison-conditions/key-facts/overcrowding/.

Correction, D. of. (2021, December 6). Accused pre-trial inmates in Correctional Facilities: Connecticut Data. State of Connecticut - Open Data. Retrieved December 7, 2021, from https://data.ct.gov/Public-Safety/Accused-Pre-Trial-Inmates-in-Correctional-Faciliti/b674-jy6w.

# 07 Appendix

- Figure 1

```
Residuals:
    Min      1Q  Median      3Q     Max
-380.64 -127.87  -28.09  118.36  562.23

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    362.786      3.463 104.747  < 2e-16
RACEASIAN                       -5.547      4.120  -1.346 0.178195
RACEBLACK                       18.858      3.476   5.425 5.80e-08
RACEHISPANIC                    10.239      3.482   2.941 0.003276
RACEWHITE                      -49.214      3.482 -14.134  < 2e-16
OFFENSENEWM                   -109.048      8.436 -12.927  < 2e-16
OFFENSENEWOther                -57.220      7.421  -7.711 1.25e-14
RACEASIAN:OFFENSENEWM          -91.107     10.565  -8.623  < 2e-16
RACEBLACK:OFFENSENEWM          -93.206      8.475 -10.997  < 2e-16
RACEHISPANIC:OFFENSENEWM       -84.096      8.491  -9.904  < 2e-16
RACEWHITE:OFFENSENEWM          -29.429      8.471  -3.474 0.000512
RACEASIAN:OFFENSENEWOther       21.640      8.916   2.427 0.015221
RACEBLACK:OFFENSENEWOther      -12.256      7.447  -1.646 0.099816
RACEHISPANIC:OFFENSENEWOther   -21.772      7.460  -2.918 0.003518
RACEWHITE:OFFENSENEWOther       -4.056      7.449  -0.544 0.586116
```

- Figure 2

```
                               Estimate Std. Error  t value
(Intercept)                   4.006e+02  1.669e+02    2.400
LATEST.ADMISSION.DATE        -1.793e-02  4.328e-04  -41.436
RACEASIAN                    -1.086e-01  4.048e+00   -0.027
RACEBLACK                     1.251e+01  3.414e+00    3.665
RACEHISPANIC                  5.902e+00  3.420e+00    1.726
RACEWHITE                    -4.113e+01  3.420e+00  -12.027
GENDERM                       6.452e+01  5.024e-01  128.425
AGE                          -1.740e+00  1.344e-02 -129.505
DETAINERDO NOT RELEASE        1.656e+02  1.925e+02    0.860
DETAINERFEDERAL               3.853e+02  1.667e+02    2.312
DETAINERGOVERNOR WRNT         5.694e+02  1.800e+02    3.162
DETAINERH                     2.142e+02  1.680e+02    1.275
DETAINERIMMIGRATION           2.738e+02  1.667e+02    1.642
DETAINERNONE                  2.880e+02  1.667e+02    1.728
DETAINEROTHER STATE           3.088e+02  1.667e+02    1.852
DETAINERSPECIAL PAROLE        1.902e+02  1.667e+02    1.141
DETAINERSTATE OF CT           3.080e+02  1.667e+02    1.848
DETAINERT                     3.564e+02  1.667e+02    2.138
OFFENSENEWM                  -1.059e+02  8.284e+00  -12.785
OFFENSENEWOther              -7.413e+01  7.288e+00  -10.172
RACEASIAN:OFFENSENEWM        -9.202e+01  1.038e+01   -8.869
RACEBLACK:OFFENSENEWM        -8.481e+01  8.323e+00  -10.190
RACEHISPANIC:OFFENSENEWM     -8.074e+01  8.339e+00   -9.683
RACEWHITE:OFFENSENEWM        -2.675e+01  8.318e+00   -3.216
RACEASIAN:OFFENSENEWOther     3.064e+01  8.758e+00    3.499
RACEBLACK:OFFENSENEWOther     7.140e+00  7.314e+00    0.976
RACEHISPANIC:OFFENSENEWOther -1.927e+00  7.327e+00   -0.263
```

- Figure 3

| K | RMSE |
|---|---|
| 1 | 1035.3917 |
| 3 | 586.8484 |
| 5 | 698.6282 |
| 7 | 744.7195 |
| 9 | 807.9662 |
| 11 | 817.0356 |