Who is ready to leave (Criminal version)

Yang Chang, Dongting Ma

Abstract

Across the nation, judges, probation and parole officers are increasingly using algorithms to assess a criminal defendant's likelihood of becoming a recidivist. A tool called COMPAS is introduced and used in many jurisdictions around the U.S. to predict if a convicted criminal is likely to re-offend. COMPAS algorithms assign a score to each defendant ranging from 1 to 10 with ten being the highest risk.With quantifiable data, we could potentially build a model to predict if a criminal defendant will commit crime again or not. Therefore, Our goal for this project is to train a model using crime history data, predict criminal defendants' likelihood of becoming recidivists and thus help decide which inmates are ready for parole.

1. Data background

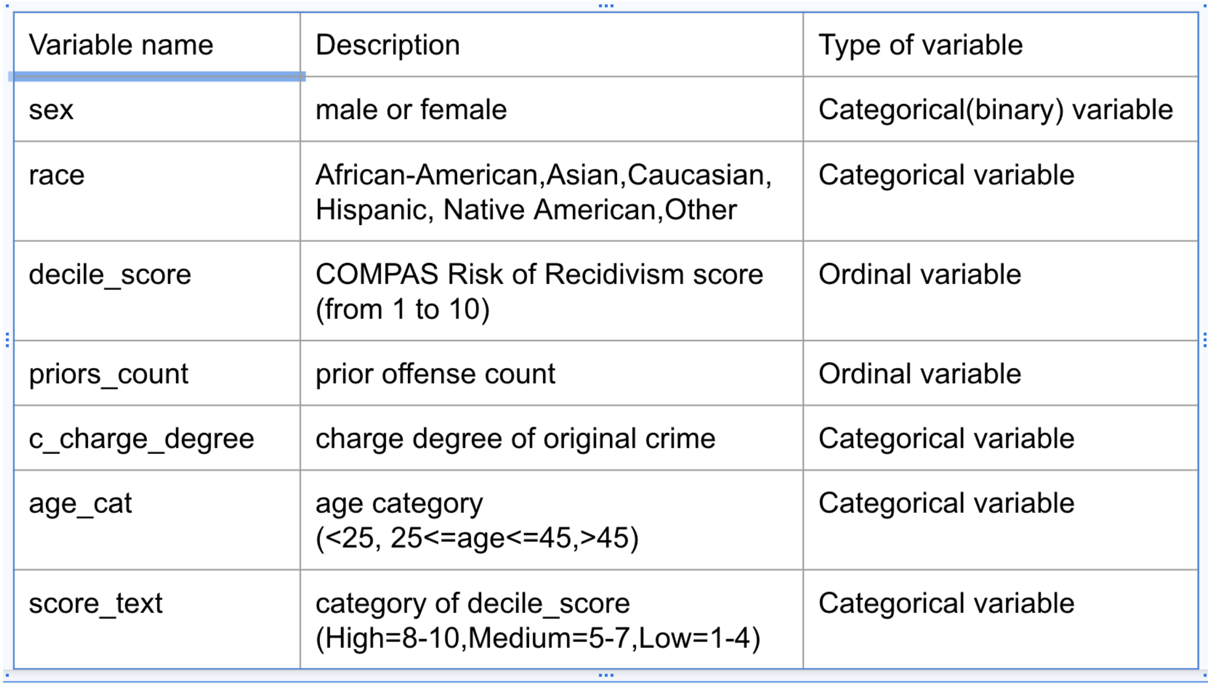
We looked at more than 7,000 criminal defendants data obtained two years' worth of COMPAS scores from the Broward County Sheriff's Office in Florida and compared the predicted recidivism rates with the rate that actually occurred over a two-year period. These 7,000 criminal defendants' age range from 18 to 96 with a mean age of 34. Most defendants are booked in jail where they respond to a COMPAS questionnaire. Their answers are fed into the COMPAS software to generate several scores including predictions of "Risk of Recidivism" and "Risk of Violent Recidivism."

2. Messy data and data cleaning

There are over 50 predictor variables in the raw dataset. However, there are over fifteen variables which have more than 50% missing value. For those variables, it is really hard to fill them so we decide to drop them. Furthermore, we also find repeat columns which have exactly the same data in them. We choose to keep one and drop another so that it will not affect training our model. There is one variable named 'days_b_screening_arrest' which means days between arrest and screening. According to a study conducted earlier about COMPAS score , if it is over 30 days, it has low data quality.So we filter observations with an absolute value of ‘days_b_screening_arrest’ more than 30 . Lastly, We also drop variables that have only one category in them because it is not helpful when fitting our model.

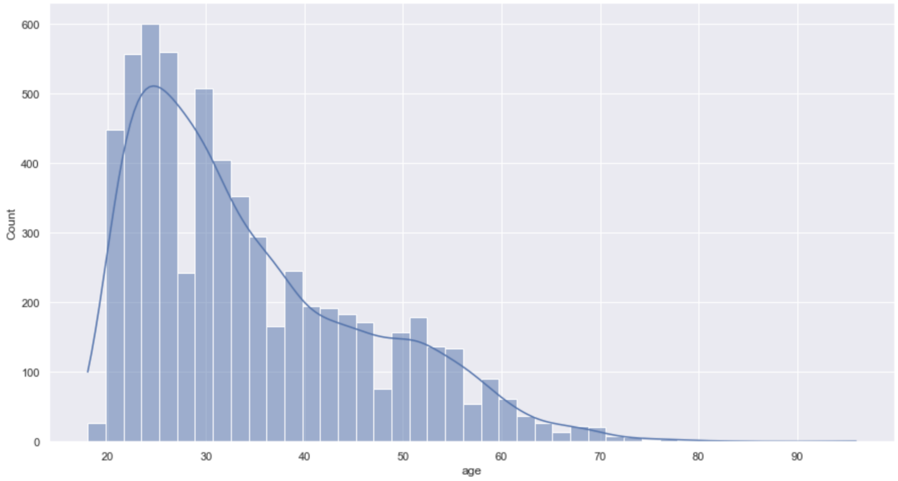
3. Data visualization

List of important variables:

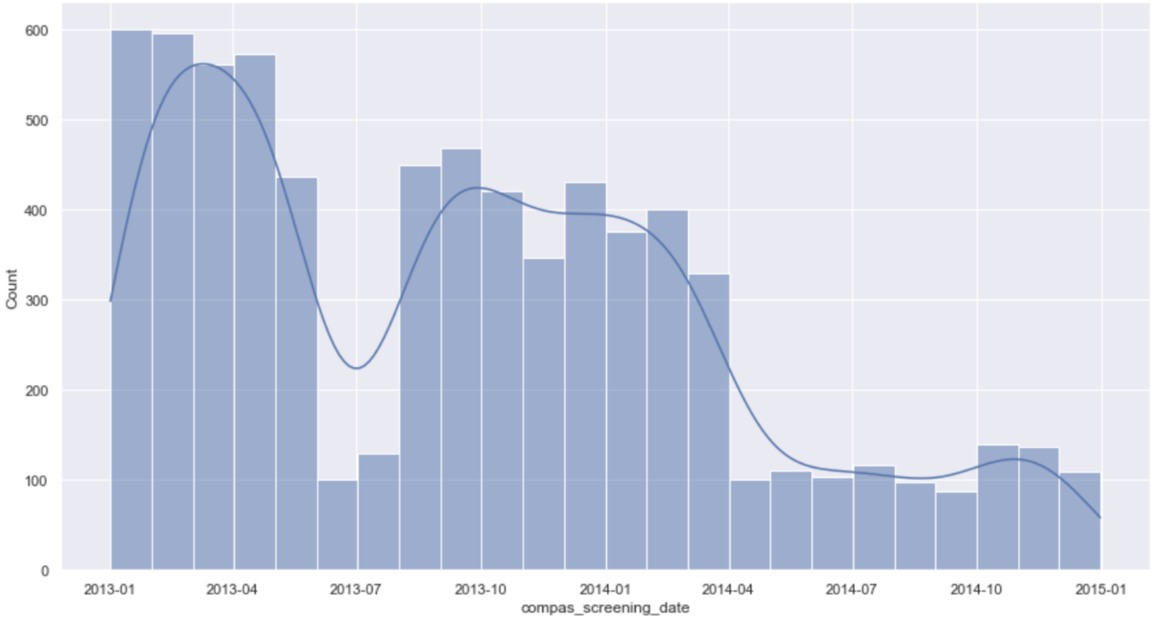


Variable name	Description	Type of variable
sex	male or female	Categorical(binary) variable
race	African-American,Asian,Caucasian, Hispanic, Native American,Other	Categorical variable
decile_score	COMPAS Risk of Recidivism score (from 1 to 10)	Ordinal variable
priors_count	prior offense count	Ordinal variable
c_charge_degree	charge degree of original crime	Categorical variable
age_cat	age category (<25, 25<=age<=45,>45)	Categorical variable
score_text	category of decile_score (High=8-10,Medium=5-7,Low=1-4)	Categorical variable

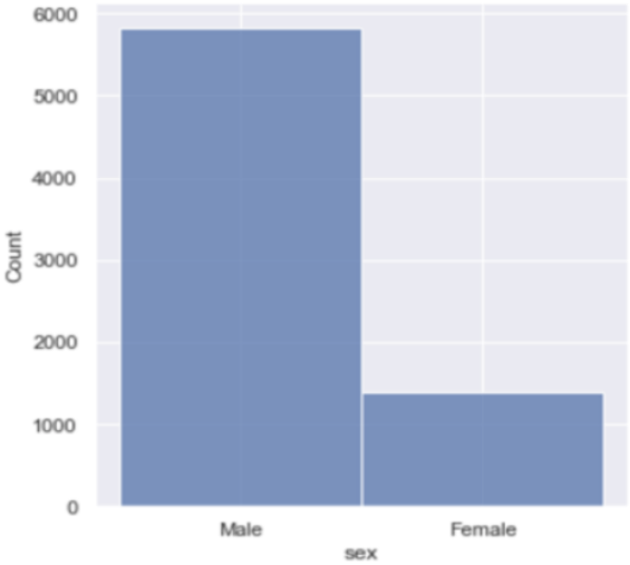
We include a list of important variables here for better interpretation of our project and all of them are chosen to fit our model later.



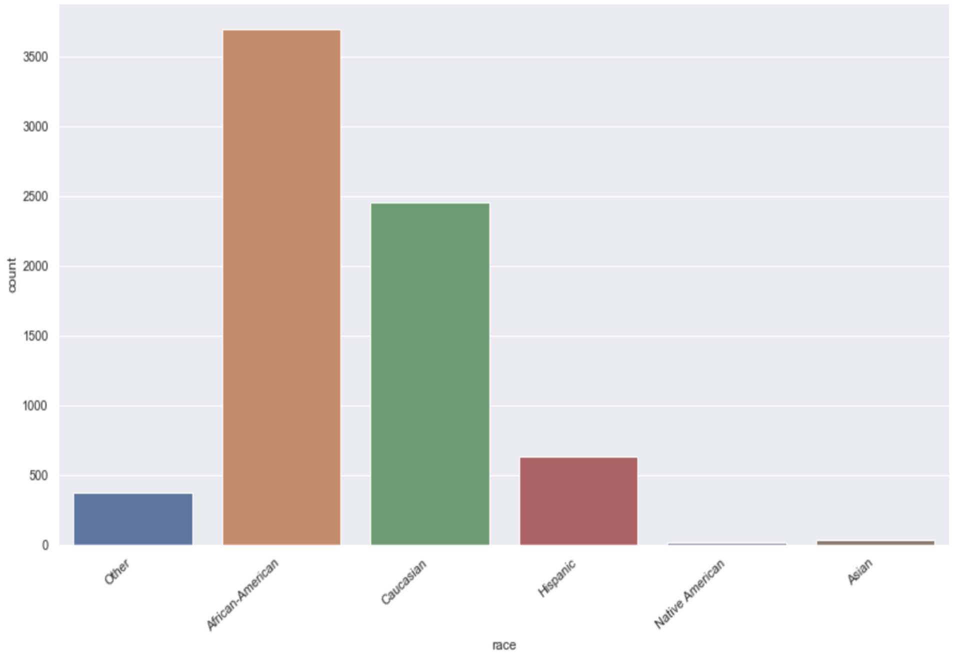
We discover the age mainly lies in the 20 to 35 age group with right skewed distribution. The minimum age is 18 and no defendants are under the age of 18 which indicates no data error.



We also did a time-series analysis and we noticed that the COMPAS screening for 2013-07 and after 2014-04 are fewer for some reason. We think it is because the data was obtained in 2016 and we are assessing the recidivism rate over a two-year period so less data is available now till the two-year period reaches.



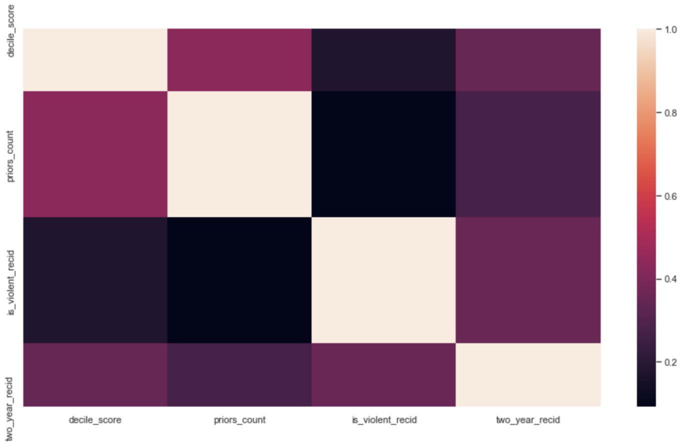
The number of male criminal defendants is almost three times more than the number of female criminal defendants



. The race composition of criminal defendants is consistent with Broward County racial demographics. The majority race in the sample is African-American, followed by Caucasian, and Hispanic.

4.Feature engineering

As we still have more than 30 features after data cleaning, we need to choose features by importance and best for our model. We decided to use backward selection to select the best 15 features. In the result, we found features that are exclusive to every observation such as name and date of birth. We think these two variables are not useful when training our model so we just drop them.



With final features chosen, we would like to see the correlation between our features to prevent multicollinearity and find the possible interaction term. As we could see here, the lighter color means more correlation between the two variables.

Furthermore, Since this dataset includes a lot of ordinal variables and nominal variables such as sex,race, score_text, and age_cat, we decide to use dummy encoding and end up getting 20 variable for model fitting.

5. model selection and model training

For this project, we would like to predict if a criminal defendant will commit crime again after paroling. We selected three classification models that we learned in the class: Logistic regression, Linear Support Vector Classifier and Random Forest Classifier. We split the data set into a 50/25/25 training/validation/test split. Training dataset is for model fitting while validation dataset is for hyperparameter tuning and test dataset is for evaluating model performance.

Logistic Regression

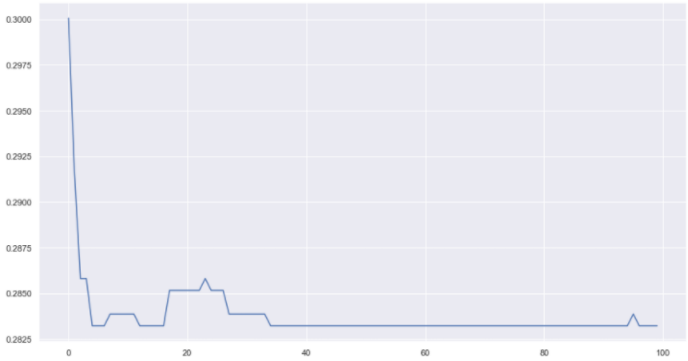
$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Logistic Regression is a good method to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables, which is really suitable for our project.

$$\min \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y; w) + \lambda \cdot ||w||_2^2$$

We first used the validation dataset to determine an appropriate regularization λ by looping through $\lambda \in \Lambda = \{.01,.02,.03,...,1\}$ to minimize the regularized empirical risk for logistic loss $(log(1 + exp(-\ yw^T x)))$ using the above formula. The graph below shows the error rate for each responsive λ.

As we can see, there are multiple λ that have the lowest error rate (0.28) on validation dataset. We selected one of them (0.04) as our best

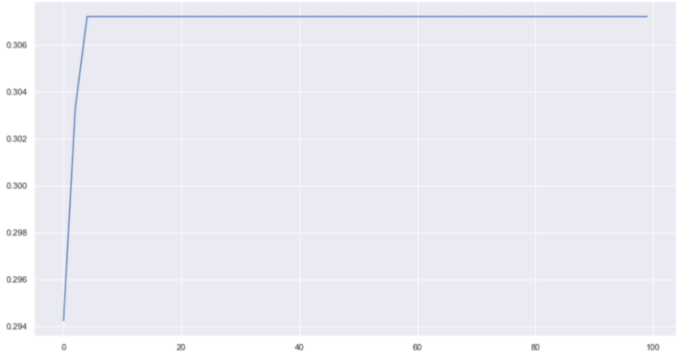


regularization parameter. Then we fitted the model using the parameter and we got an error rate of 0.28 on the test dataset.

Finally, we printed out the coefficients for each feature and found out that "is_violent_recid" tended to be a much more important feature than others when predicting who will reoffend. This made sense because she/he more likely reoffend if he/she has reoffended violently before.

Linear Support Vector Classifier

Support Vector Machine is to breast the best line or decision boundary, the hyperplane, that can segregate n-dimensional space into classes so that we can easily put the new data in the correct category in the future. We used the same way as used for logistic regression to minimize the regularized empirical risk for hinge loss ($(1 - yw^{T}x)\ +$). The graph below shows the error rate for each responsive λ.



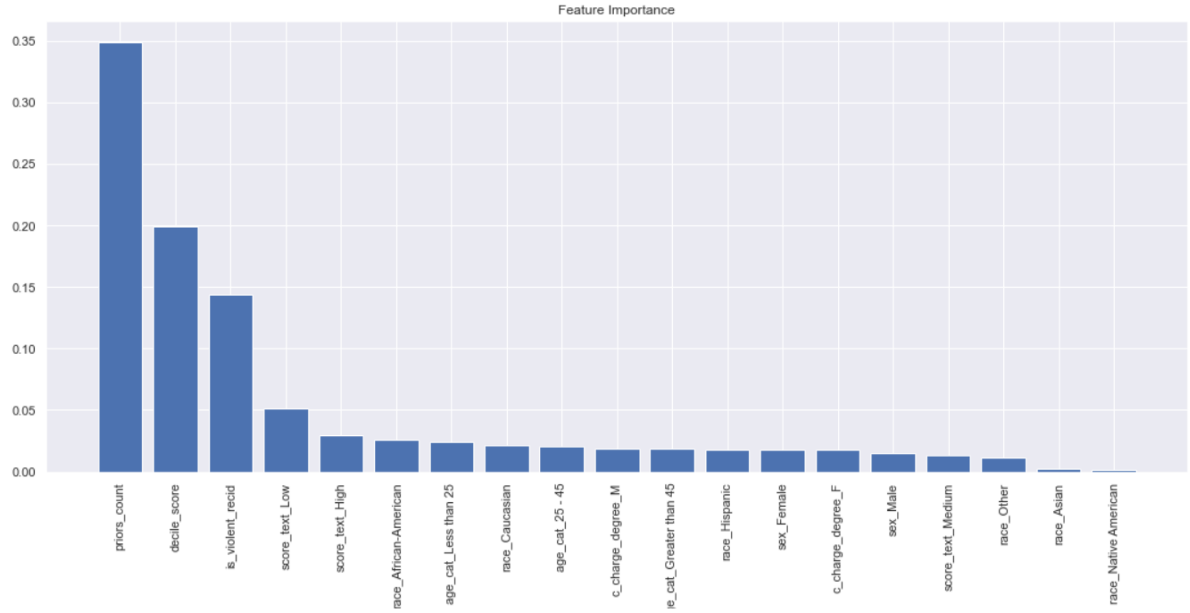
As we can see, the error rate increases as λ increases on validation dataset. So λ = 0.01 has the lowest error rate of 0.29 on validation dataset. Then we fitted the model using the parameter and we got a test error rate of 0.3.

Finally, we also printed out the coefficients for each feature. The absolute size of the coefficient relative to the other ones gives an indication of how important the feature was for the separation. We found out that "is_violent_recid", "score_text_low" and "score_text_high" had the largest absolute size of coefficients which tended to be the most important features in this model.

Random Forests Classifier

The random Forests Classifier is an ensemble learning method for classification that operates by constructing a multitude of decision trees at training. The output of the random forests classifier is the class selected by most trees. Since Linear SVM and logistic regression are both linear models, then we tried a non-linear model to do the prediction task.

We first fitted a Random Forest Classifier using default parameters on the training set. We only got a test error rate of 0.32. Then we used grid search to find the best hyperparameter. By trying 100 different combinations of each parameters on the validation dataset, we found that the best combination of parameters is n_estimators = 400, min_samples_split = 10, min_samples_leaf = 4, max_features = auto, max_depth = 70 and bootstrap = True. By fitting the model using these parameters, we successfully decreased the test error rate from 0.32 to 0.27.



Finally, we used the feature importance package of Random Forest to see which features are most relevant. Feature importance is computed as the mean and standard deviation of accumulation of the impurity decrease within each tree. According to the graph above, “Priors_count”, “Decile_score” and “is_violent_recid” are the most important features of the Random Forest Classifier.

6.Model Comparison and Evaluation

Model	Recall Rate	Precision Rate	F1	Error Rate	Accuracy
Logistic Regression	0.6	0.73	0.66	0.28	0.72
LinearSVC	0.58	0.72	0.64	0.30	0.7
RandomForest Classifier	0.63	0.74	0.68	0.27	0.73

The table above shows comparison of different evaluation metrics among three models. The RandomForest Classifier outperformed the other two models for all evaluation metrics. Accuracy and F1 reached 0.73 and 0.68, respectively. The reason why RandomForest performs better might be that there is not a strong linear relationship between dependent and independent variables. We also noticed that the precision rate is higher than the recall rate in general which means that the model returns not many results but most of the results are correct compared to the training label.

7.Conclusion and Future Move

By implementing three models and comparing their performance, we found out that the Random Forest Classifier performed best and reached 0.73 and 0.68 for accuracy and F1, respectively. “Priors_count”, “Decile_score” and “is_violent_recid" were considered the most important features by using this model while “is_violent_recid” was considered as the most important features among all three models. We believe that this model is a so-called Weapon of Math Destruction since the predictions can have negative consequences and can create self-fulfilling feedback loops. If the model mislabeled a prisoner as someone ready to leave which he/she is not, he/she will reoffend and harm the society. Also, the bias toward a specific racial group of the prediction may cause racial hatred. Furthermore, keeping prisoners in prison longer may cause them to reoffend. People labeled as “likely to reoffend” will be kept longer and it will be hard for them to find a job when they leave. They will likely break laws for living such as robbery, burgary and even more serious crimes. Thus, a negative self-fulfilling feedback loop is created.

We also evaluated the fairness of this model among different races. We discovered a bias towards African Americans. African Americans are more likely to be predicted as “likely to recidivate”. There are 596 people who are predicted to be a recidivist. However, 68% (398) of them are African Americans. Also, the feature importance graph shows a much higher importance of “race_African_American” than other races. Racial fairness of this model is important since otherwise it will cause racial hatred. However, it is very hard to define what is considered as fair and how to improve it. The reason why the model biases towards African Americans is because there is a high proportion of African American prisoners (52%) in the dataset.

Thus, we think that the model is not ready to be used in production to determine who is ready to leave since it does not have a very high accuracy as well as F1 and it biases towards African Americans. Here are some suggested future moves. Firstly, to improve the fairness, records of other races can be duplicated to reduce the high proportion of African Americans. More evenly distributed races might help improve the fairness of the algorithm. Secondly, implementing other more complex models such as neural networks might improve the accuracy. Thirdly, as the data from the Broward County Sheriff’s Office should become more every two year, we could potentially train our model with more data. As time goes by, the data quality will increase and less missing could bring more features available for us to train our model.

Reference:

Julia Angwin, Jeff Larson. “Machine Bias.” *ProPublica*, 23 May 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Compas analysis, thejefflarson,
https://github.com/propublica/compas-analysis/