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. Our goal for this project is to train a model using crime history data, predict criminal defendant's likelihood of becoming a recidivist and help decide which inmates are ready for parole.

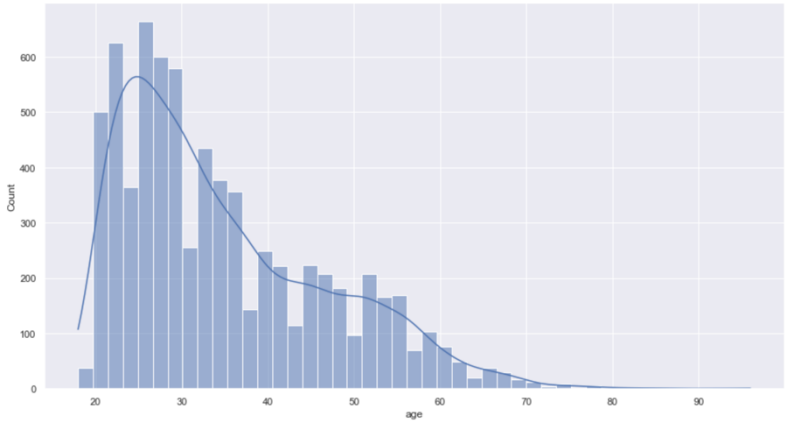
1. Data background

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 their 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, 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."

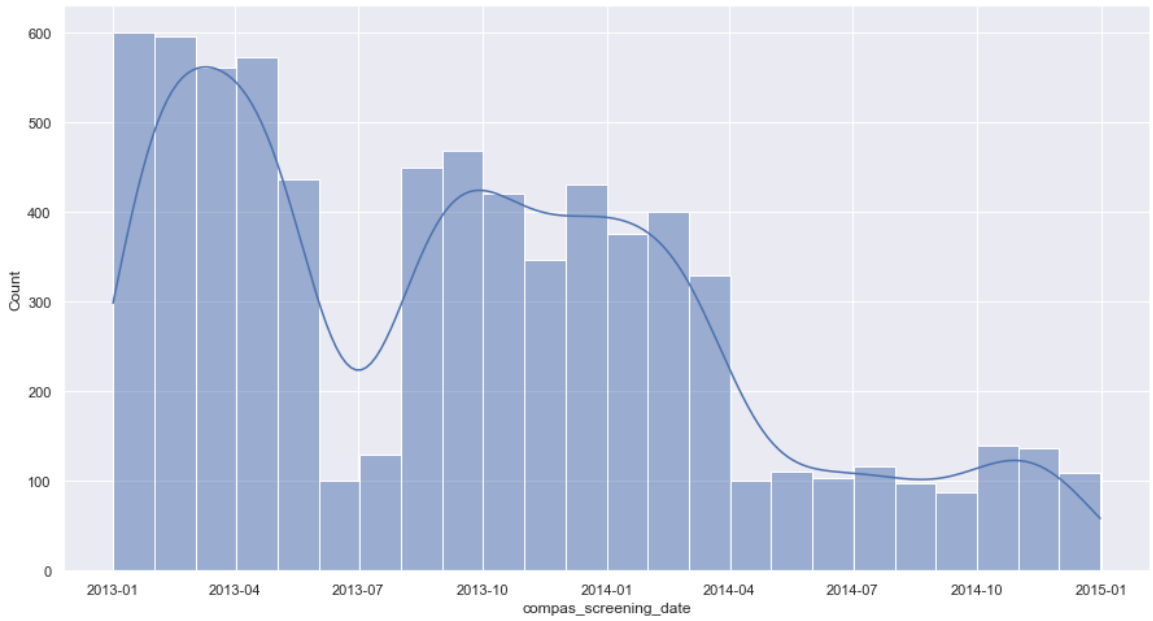
2. Messy data and data cleaning

There are over 50 predictor variables in the raw dataset. As the data is created by government authorities, most data are complete and only two variables have missing data. One is named "r_days_from_arrest". This variable means number of days between follow-up crime and arrest date. If the criminal defendants do not commit any follow-up crime, this value should be NA. Therefore, we decide to deal it with dummy encoding. Another variable with missing data is called "violent_recid" and there is nothing in this variable. We just simply decide to drop this variable since it is not helpful at all. We also find negative days data which is input by mistake. We decide to drop error data since they are only a few of them.

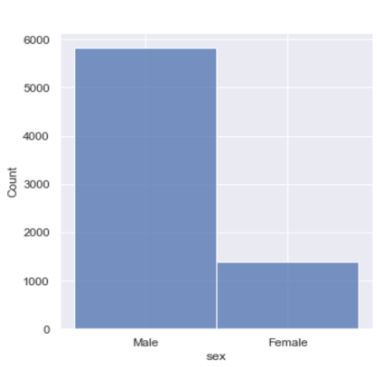
3. Data visualization

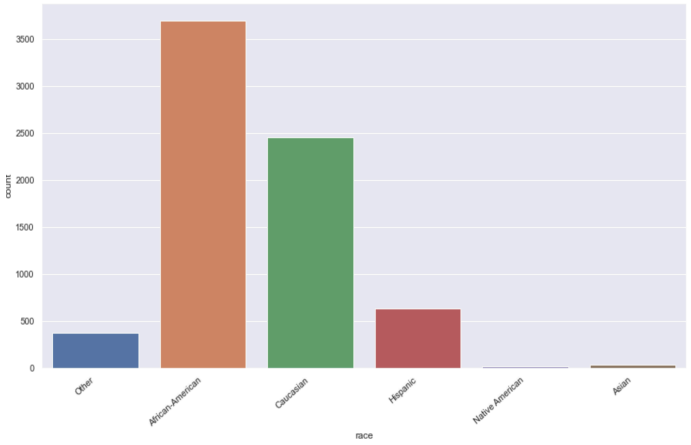


We discover the age mainly lies in the 20 to 35 age group with right skewed distribution. The minimum age is 18 and no defendants are under the age of 18 which indicates no data error.

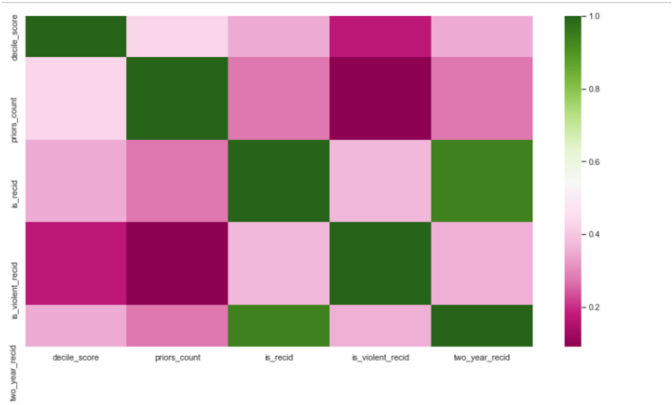


We also did a time-series analysis and we noticed that the compas screening for 2013-07 and after 2014-04 are fewer for some reason. We are not sure if it is missing data, or something happened in those periods. we will discover the reason in the future.





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.



We briefly choose some features that may be good for training our model and we would like to determine how correlated our different features were to one another. We will add more features as we get along with fitting our model. From the correlation graph on the left, we can see that variable in green have a large correlation whereas variable in dark pink have a small correlation

4.Dummy encoding and feature scaling

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 variable to be necessary before doing training our model. As we decide to use tree model to train our model, feature scaling is not a must for tree model.

5. model selection and model train

```
rdc=RandomForestClassifier(n_estimators=100)
rdc.fit(X,Y)
cross_val_score(rdc, X, Y,
                scoring="neg_mean_squared_error",
                cv=3).mean()
```

```
/opt/anaconda3/lib/python3.7/site-packages/ipykernel_launch
ed when a 1d array was expected. Please change the shape of

/opt/anaconda3/lib/python3.7/site-packages/sklearn/model_se
n-vector y was passed when a 1d array was expected. Please
avel().
  estimator.fit(X_train, y_train, **fit_params)
/opt/anaconda3/lib/python3.7/site-packages/sklearn/model_se
n-vector y was passed when a 1d array was expected. Please
avel().
  estimator.fit(X_train, y_train, **fit_params)
/opt/anaconda3/lib/python3.7/site-packages/sklearn/model_se
n-vector y was passed when a 1d array was expected. Please
avel().
  estimator.fit(X_train, y_train, **fit_params)
```

```
-0.3092596308070518
```

For this project, we would like to predict if a criminal defendant will commit crime again after paroling. Since our outcome is binary, we think random forest is a good start. Without only partial features and no model optimization, we do not expect a perfect result. We only get a negative mean squared error of -0.30 which means the MSE is 0.3. We will keep adding more highly relevant features and drop irrelevant ones when building our model.

6. Midterm conclusion

Overall, I think we have a good start. We have a complete interpretation of all the features in the dataset which is helpful for our feature engineering. As for the remaining time of this semester, we plan to spend more time on feature engineering and model optimization to reduce the error of our model. We hope we can get a high prediction accuracy with our final model.