2020

University of South Florida
ISM 6905.001S20
Independent Study Project

ANALYZING MACHINE BIAS IN COMPAS RISK ASSESSMENT SYSTEM

GUIDE : PROF. ANOL BHATTACHARJEE, PROF. JONI JONES

AUTHOR: RUSHIKESH MANOHAR MAHESHWARI

U41912945

Contents

Introduction 2

Literature review 2

Data 3

Objective 3

Method 3

Data pre-processing and Feature Engineering 4

Analysis and Results 5

1. Building Fair Risk Score 5
2. Building fair_score using crime related factors 6
3. Predicting Recidivism with Fair and Decile score 8

Future Scope: 9

Reference: 9

Introduction

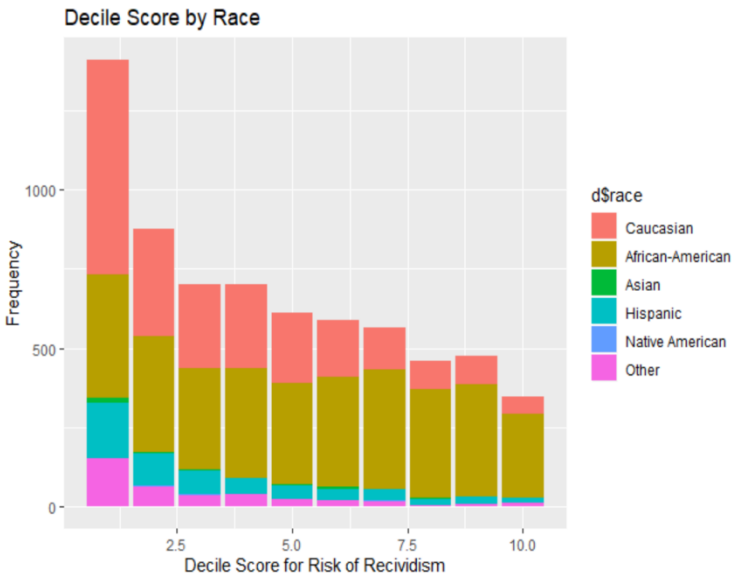
Risk assessment algorithms and biased results are not pleasant combination. In this paper we revisit the famous 'Machine Bias' article published by ProPublica in May 2016 about how the COMPAS systems used for pre-trial risk assessment is bias towards a race. Our work questions and corrects rudimentary assumption ProPublica made which turns out to be an error. We also make an effort to build a fair assessment score for the given data only based on criminal activities, eliminating the race or any racial effect in the outcome.

Literature review

United States has the world's highest incarceration rate of 655 per 100,000 of population as per the world prison brief data. To control this rate and inflow of the prisoners, states of New York, Wisconsin, California, Florida and few other jurisdictions uses **COMPAS** (developed and owned by Northpointe), an acronym for *Correctional Offender Management Profiling for Alternative Sanctions* which provides certain risk score for an offender to access its likelihood of being a recidivist. These scores are used by US courts to grant the pre-trial release too.

ProPublica, an American non-profit newsroom based out of New York, presented a study in May 2016 and claimed that COMPAS generated 'Risk Scores' are biased against African-American race group. For this study they looked at more than 10,000 criminal defendants in Broward County, Florida, and compared their predicted recidivism rates with the rate that actually occurred over a two-year period. To do so, they also collected data about public incarceration from the 'Florida Department of Corrections'. By joining these two data sets with a defendant's first name, last name and Date of birth ProPublica assess about 11000 records.

Their analysis was at fault in terms of sampling error and terminologies as they swapped the cause-effect which inflated the result. The article '*False Positives, False Negatives, and False Analyses' : A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks.",* states how ProPublica has failed to test the bias as per standard and used faulty statistical references and benchmarks to produce the result.



Moreover recently, ProPublica's COMPAS data was revisited by Matias Barenstein, published on July 08, 2019. He highlights a classic data sampling mistakes made by ProPublica. In which ProPublica included extra 40% (1000) recidivists in the analysis which inflated the results of racial bias by 24%. In order to check for recidivism, ProPublica should have used a cut off on the COMPAS screening date (at 1st April 2014). This would ensure that these people are observed for next two years, as the database contains data through March 2016.

Data

Objective

We studied and recreated ProPublica's analysis and based on that study, we defined 3 modes of analysis:

- To study the marginal effects of available attributes in data on the decile score (risk score) generated using COMPAS which is used by Courts to tag an offender with 'low', 'medium', or 'high' risk of recidivism
- To build 'fair score' from the available data that only concerns to the criminal history and current crime description of an offender and no other social factors, and compare its variation with original 'decile score'
- To compare the predictive power and error rate of 'decile score' and 'fair decile score' for recidivism

Method

ProPublica provided a copy of database they worked with on their GitHub repository. They had acquired this dataset from Broward County, FL. It includes the pretrial defendant's data from January 2013 to December 2014. They also got the data from Broward county's Sherriff's office and federal jail to observe these defendants for recidivism till April 2016. We sourced the database copy and extracted various tables which had information collected from Broward county COMPAS system.

Firstly, we replicated the data pre-processing and model built by ProPublica. We notice there are 2 errors in analysis done by ProPublica. A data dictionary is provided in annexure.

1. Sampling Bias

ProPublica selected just the recidivist population (over 1000 more records) but removed the corresponding non- recidivist population for given timeframe. This made the sample used for analysis bias by having more recidivist population and hence the results were bound to inflate.

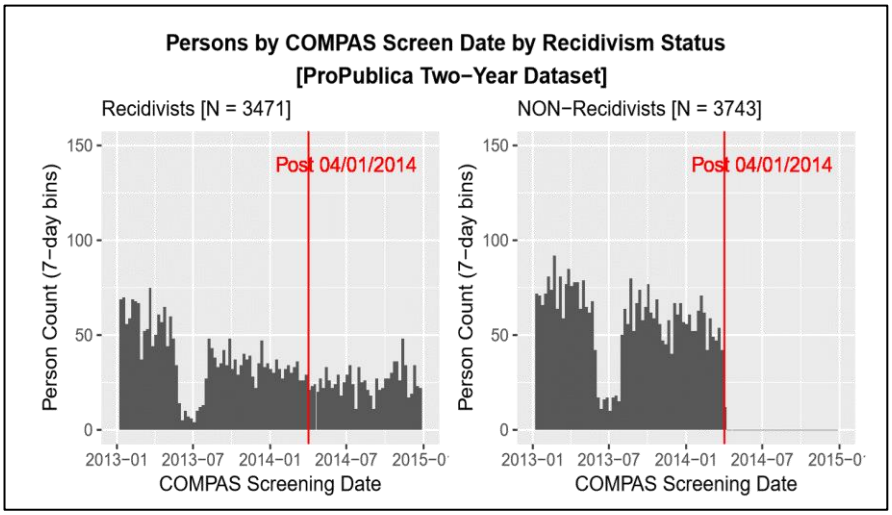


Figure 1.Segregation of defendants as recidivist and non-recidivist across compass screening date

2. Cause-Effect mix up

ProPublica collected data from Broward county's Sherriff's office so they could know if the defendant has committed and arrested again. This information was stored under 'Two-year Recidivism' variable in the data. This information was used from data in order to predict the of risk score for defendants. This data was not available while generating risk scores for any defendant in first place, so logically cannot be used in analysis. The act of recidivism (Effect) is used to predict the cause of recidivism (Cause).

Risk of General Recidivism Logistic Model

	Dependent variable:
	Score (Low vs Medium and High)
Female	0.221*** (0.080)
Age: Greater than 45	-1.356*** (0.099)
Age: Less than 25	1.308*** (0.076)
Black	0.477*** (0.069)
Asian	-0.254 (0.478)
Hispanic	-0.428*** (0.128)
Native American	1.394* (0.766)
Other	-0.826*** (0.162)
Number of Priors	0.269*** (0.011)
Misdemeanor	-0.311*** (0.067)
Two year Recidivism	0.686*** (0.064)
Constant	-1.526*** (0.079)
Observations	6,172
Akaike Inf. Crit.	6,192.402
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 2. ProPublica's analysis indicating usage of Two_year_recidivism for analysis of Decile score

Data pre-processing and Feature Engineering:

We used RStudio to extract data from database copy. Primarily we worked on *Compas, people, charge and casearrest* tables. Following pre-processing was done on the data.

1. We joined all these tables by making unique keys by person_id, first_name, last_name, and date_of_birth of the criminal. The charge_degree_level was sourced from *charge* table to *final_data* and encoded in orderly fashion from 1 to 6 where 1 being M3 ((misdemeanour level 3) all the way up to 6 being F1 (felony level 1)
2. Deleted all the records where recidivism flag was -1 i.e. was missing
3. Created a new feature of 'drug_involvment' from the charge_description by collecting list of drugs like (cocaine, cannbies, meth) from the charge description of a criminal and flagged criminal as involved in drug related crime. The hypothesis here is criminals involved in drug are more likely to commit crime again due to drug addiction
4. Computed length_of_stay in a jail from jail_in and jail_out time for a criminal
5. Only considered records till time 1st April 2014 and removed the sampling bias pro-publica has in their analysis

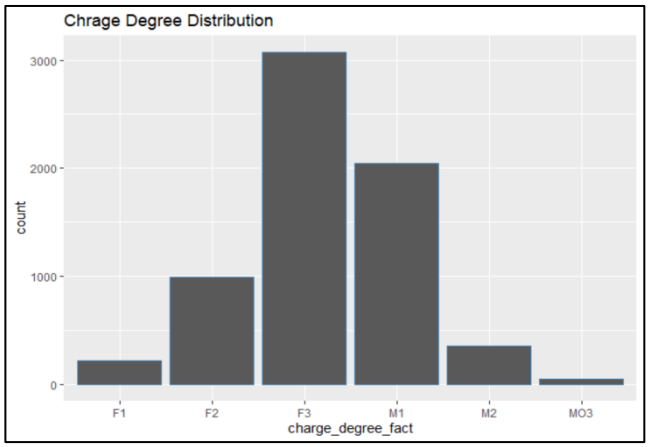


Figure 3. Distribution of Charge Degree

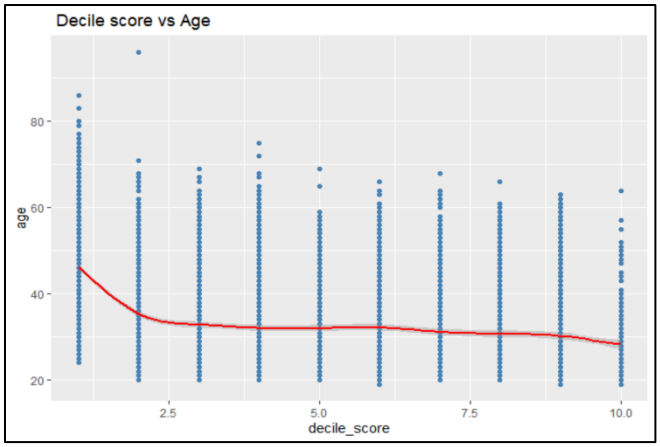


Figure 4. Relation between Decile score and Age of criminal

Analysis and Results:

As stated in the objective, our analysis was aimed to build a fair risk score which is not racially bias and has which is also a good predictor of act of recidivism.

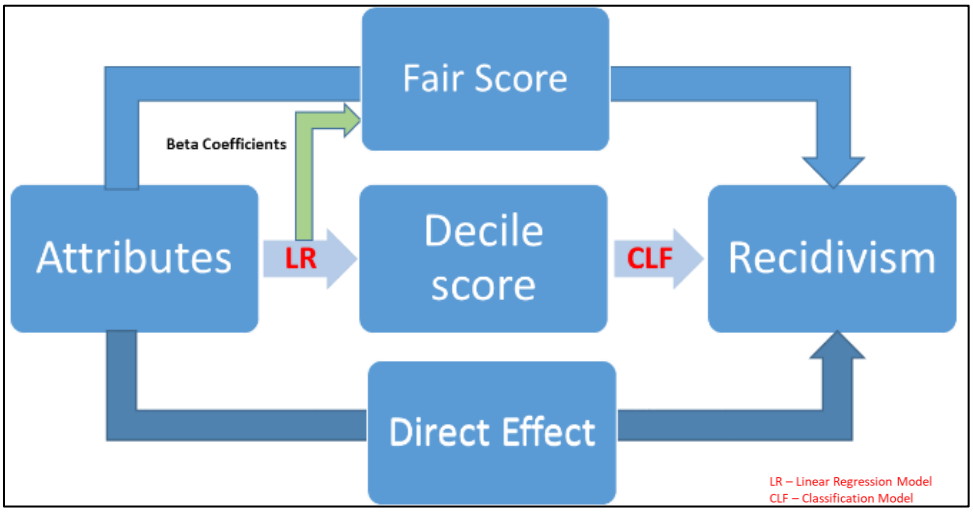


Figure 5. Methodology used for building fair score and predicting act of recidivism

1. Building Fair Risk Score

Since the claim is that the decile score provided by COMPAS system is biased towards African-American population, we intended to build a fair score which would not account for any racial factor. So, we studied the individual and combined effects of attributes on decile score. Since the decile score was highly right skewed we used logarithmic transformation to convert it into normal distribution to perform linear regression.

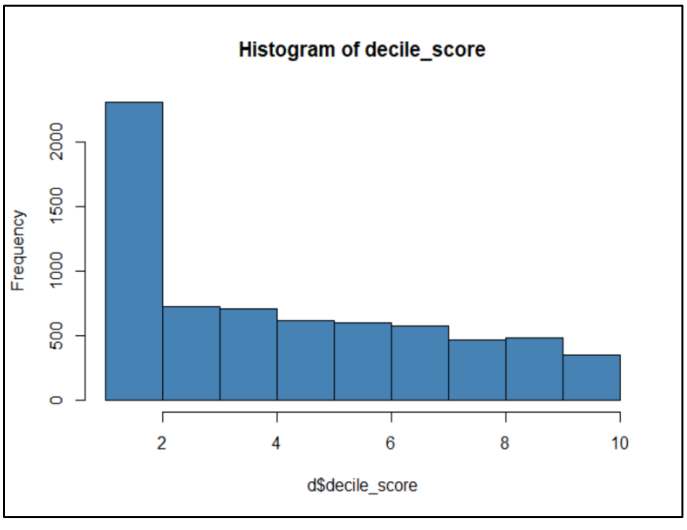


Figure 6. Distribution of Decile score

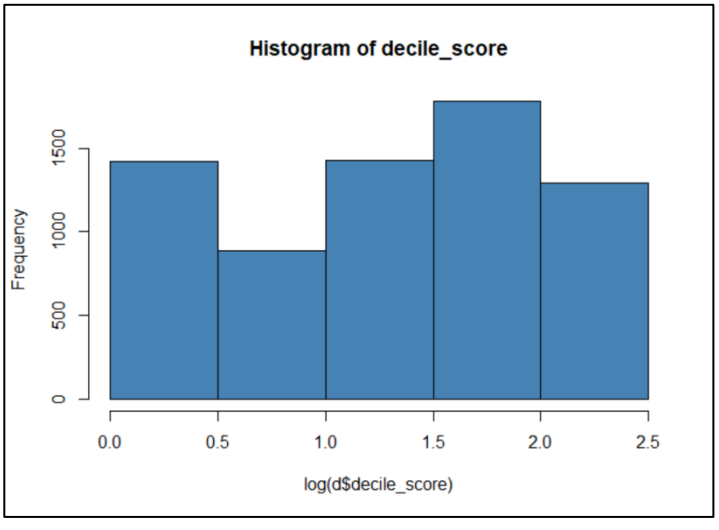


Figure 7. Log Transformed Distribution of Decile score

1. Decile score as a function of crime and demographic factors

$$Decile\ score = \beta_0 + \beta_1(age) + \beta_2(Juv\_fel\_count) + \beta_3(Juv\_misd\_count) + \beta_4(sex) + \beta_5(priors\_count) + \beta_6(race) + \beta_7(charge\_degree) + error$$

2. Decile score as a function of crime and demographic factors with sex and race interaction

$$Decile\ score = \beta_0 + \beta_1(age) + \beta_2(Juv\_fel\_count) + \beta_3(Juv\_misd\_count) + \beta_4(Sex * race) + \beta_5(priors_{count}) + \beta_6(charge\_degree) + error$$

3. Decile score as function of crime related factors

$$Decile\ score = \beta_0 + \beta_1(Juv\_fel\_count) + \beta_2(Juv\_misd\_count) + \beta_3(priors\_count) + \beta_3(charge\_degree) + \beta_4(drug\_involvment) + error$$

4. Decile score as a function of just Race and sex interaction term

$$Decile\ score = \beta_0 + \beta_1(race * sex) + error$$

We defined fair score as

'A risk score which only accounts for the crime related attributes and activities of a criminal and doesn't take into account any other demographic or socio-economic factor into consideration for the act of recidivism'

Hence we selected equation 3, Decile score as a function of crime related factors. After regressing decile score with crime related factors, we anticipate, all the racial bias is captured in the error term and hence the new fair score we built using decile score beta coefficients will be independent of racial bias

2. Building fair_score using crime related factors

In the above section where we analyzed various combination of the decile score, we particularly chose one which just related to the offender's criminal data available with us. i.e. *priors count, juv_fel_count, juv_misd_count, charge_degree, drug_involvment.* So the fair score was computed as below:

$$\begin{aligned}\textbf{Fair score} &= \exp(\beta_0) + \exp(\beta_1)(\text{Juv\_fel\_count}) + \exp(\beta_2)(Juv\_misc\_count) + \exp(\beta_3)(priors\_count) + \exp(\beta_4)(\text{charge\_degree}) \\ &= \exp(0.9611) + \exp(0.1252)(Juv\_fel\_count) + \exp(0.1051)(Juv\_misc\_count) + \exp(0.057)(priors\_count) + \\ &\quad \exp(\text{variants of charge degree with base F1})(\text{charge\_degree}) + \exp(0.2181)(druginvolvment)\end{aligned}$$

After computing fair score it was rescaled from 1 to 10 so that we can directly compare it with decile score. Figure 8. Below shows the difference in the distribution. It turns out the fair score is more densely distributed on the lower score levels (from 1 to 5) and very few criminals have actual risk score of 10 (High Risk) as compared to decile score distribution.

We also took a subset of criminal based on races and check specifically for African American population of the criminals. The figure 9 illustrate the fair score and decile score distribution while figure 10 illustrate the same comparison for Caucasian race population of criminals.

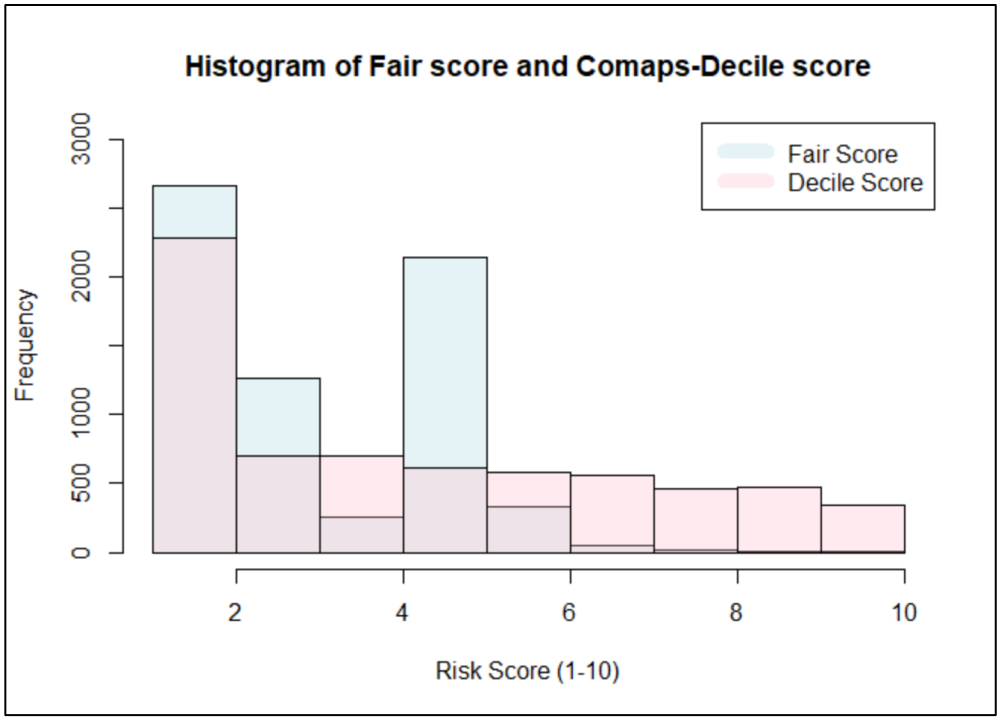


Figure 8. Fair score and Decile score distribution

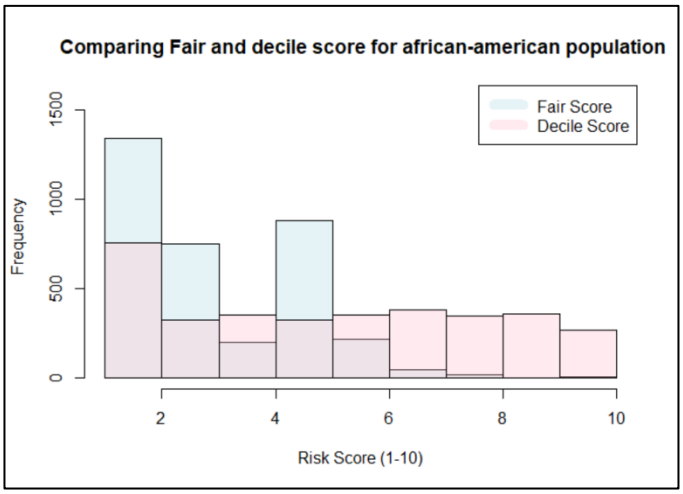


Figure 9. Fair and decile score for African-American criminals

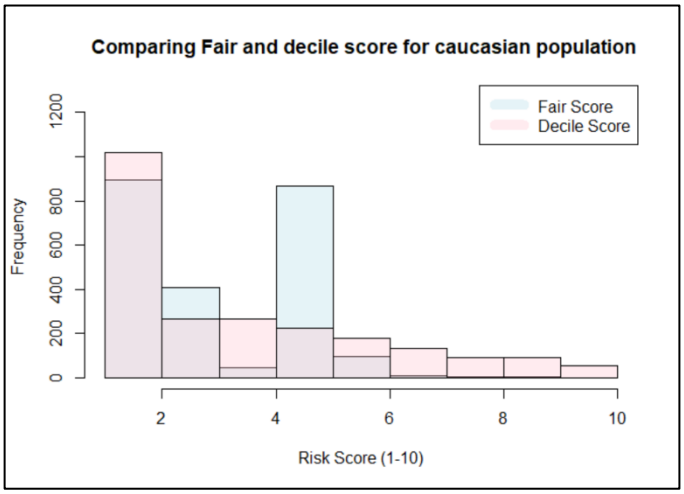


Figure 10. Fair and decile score for Caucasian population

We can see from above figures that the overall distribution of fair score for african american population is less towards the highers risk scores. The table below indicates the percent change in the average risk score of african american population is 10 times more comapred to Caucasian population of criminals.

Content	Average Decile score	Average Fair score	Percent change
Entire Criminal Population	4.458	3.451	-22.58%
African american Population	5.363	3.421	-36.21%
Caucasian (white) Population	3.622	3.499	-3.03%

3. Predicting Recidivism with Fair and Decile score

After computing fair score, now we wanted to check if decile score or fair score both has any predictive power towards recidivism. So we used both these scores separately to predict the recidivism for criminals which is a binary classification model using logistic regression.

1. Recidivism as a function of Decile score

   $Recidivism = \beta_0 + \beta_1(Decile\_score) + error$

2. Recidivism as a function of Fair score

   $Recidivism = \beta_0 + \beta_1(Fair\_score) + error$

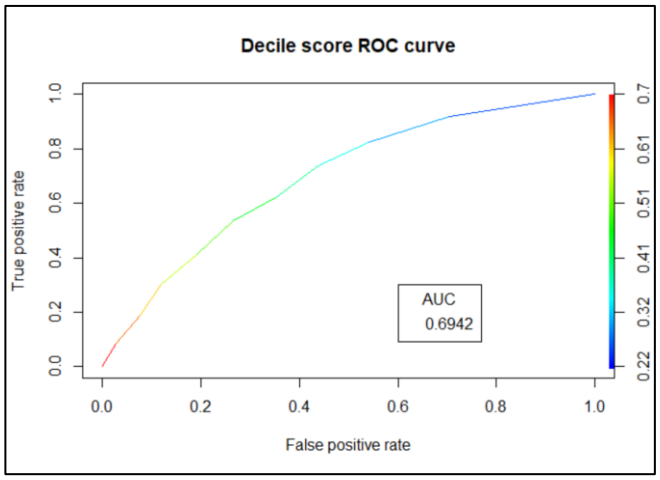


Figure 11. ROC curve for Recidivism by Decile score

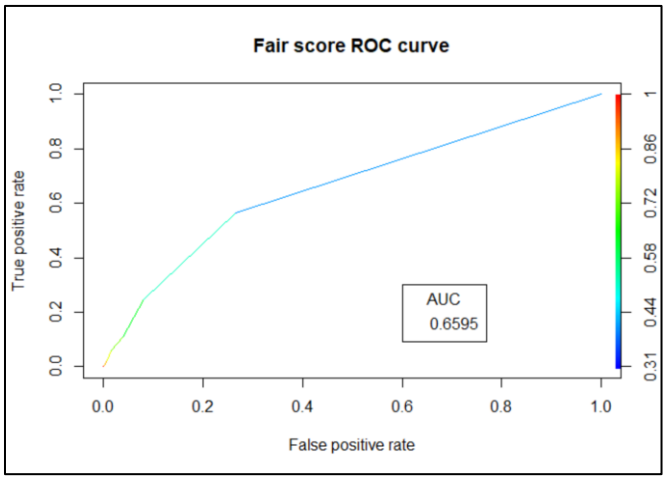


Figure 12. ROC curve for Recidivism by Fair score

Evaluation Criteria	Decile Score	Fair score
Accuracy	4.458	3.451
Sensetivity (Recall)	72%	66%
Precision	74%	91%
F1-score	0.73	0.77
AUC	0.69	0.65

Based on above result table, it can be stated that at lower cut off values, Fair score is a good predictor of recidivism but overall decile score is still a better predictor of recidivism compared to fair score as AUC for fair score is slightly lesser than that of decile score.

Future Scope:

From analysis, we can state that although the fair score cannot predict recidivism as good as decile score, looking at the percentage changes in the risk score, we can say fair score is not biased towards any race and provides rational risk scores for criminals as we could compute from available data. To make this analysis more realistic we can do following things:

1. Source more features and learn the process of computing decile score.
2. The errors terms we generated while regressing decile score against crime factor might as well have some explanatory power. We can clearly define the degree of racial bias if we have more features around socio-economic status of a criminal
3. We can check the direct effects of the attributes for computing the risk of recidivism. That can give us a clear picture weather any risk score is truly a good predictor of act of recidivism or not
4. As studied from literature review, many criminals undergo several programs in the prison so that they can have better quality of life once they get out of prison. Data regarding that can really help us understand the act of recidivism and factors affecting it
5. The influence of inmates can drive a person towards recidivism. It can also be included in future analysis.

Reference:

1. Machine Bias [https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing] by *Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica* May 23, 2016

2. How We Analyzed the COMPAS Recidivism Algorithm [https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm] by *Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin*

3. EVALUATING THE PREDICTIVE VALIDITY OF THE COMPAS RISK AND NEEDS ASSESSMENT SYSTEM by *tim brennan william dieterich beate ehret* Northpointe Institute for Public Management Inc.

4. False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks by *Anthony W. Flores California State University, Bakersfield* and *Kristin Bechtel Crime & Justice Institute at CRJ*

5. ProPublica's COMPAS Data Revisited *by Matias Barenstein July 08, 2019*

6. VALIDATION OF THE COMPAS RISK ASSESSMENT CLASSIFICATION INSTRUMENT by Broward Sheriff's Office Department of Community Control, Thomas Blomberg William Bales Karen Mann