Project Brief

Project title	**Predicting Recidivism applying IRAC method**
Module Name	**NICF Statistical Thinking for Data Science and Analytics(SF)**
Qualification Name	**NICF Diploma in Infocomm Technology (Data)**

Index

1. Purpose of this Project

This Project is used for Summative Assessment of Learner in the Module '**NICF Statistical Thinking for Data Science and Analytics(SF)**' of the NICF Course 'Diploma in Infocomm Technology (Data)'

2. Project Pre-requisites

You should have completed the following activities before starting the module project:

- Viewed and understood all the e-content related to the module
- Completed all the MCQ tests related to the module
- Completed all the Assignments of the module

You should have access to the Project Brief, Project Report template and should understand how to use the templates.

You should have access to Azure Machine Learning Studio. You should have installed Anaconda, in which you can access Jupyter notebooks and Python. You should understand the number of milestones and what are the milestones to be presented for each of the Mentoring Session.

3. Project Outcomes

You should perform all the tasks in the Project Activity List and prepare the following during the project:

- Implement the project on python Juypter notebook using the dataset given to you
- Prepare a Project Report as per pre-defined template

4. Project Definition

This is a predictive maintenance project whether the criminal defendant's likelihood of becoming a recidivist – a term used to describe criminals who re-offend , You will be given more than 10,000 criminal defendants in Broward County, Florida, and compare their predicted recidivism rates with the rate that actually occurred over a two-year period.

The County Jailhouse in State Z in the United States is working on reducing recidivism, which is the tendency of a convicted criminal to reoffend. In order to accomplish their objective, the Jailhouse hired a data scientist(You) to design an algorithm to predict the likelihood of recidivism of current inmates. The data inputs for this new algorithm include variables such as age, prior convictions, race and gender. The Jailhouse will use the predictions to determine whether or not to release inmates who come up for parole, or to grant early release from their sentence.

If you were the judge deciding this case, what would you rule? Apply the IRAC method to identify the legal issue you think arises out of these facts, and any rule or policy you think is on point. Then apply the rule or policy to reach your conclusion.

Finally after the analysis using IRAC method , you have to come up with conclusion whether the white or black defendants are at higher risk of recidivism and risk of violent recidivism

5. Project Task List

You will perform the tasks in the following sequence, while performing this project:

Task 1:

Analyse the Requirements related to the Scenario and justify why IRAC is a suitable solution

Explain in a single page how IRAC meets the objective of this Project.

Task 2: Read and prepare the data for Risk of Recidivism

- Import modules needed to implement predictive maintenance, R is used (ggplot and dplyr)
- Read the non violent dataset to read the number of rows
- Remove rows based on following conditions
    - If the charge date of a defendants Compas scored crime was not within 30 days from when the person was arrested, we assume that because of data quality reasons, that we do not have the right offense.
    - We coded the recidivist flag -- is_recid -- to be -1 if we could not find a compas case at all.
    - In a similar vein, ordinary traffic offenses -- those with a c_charge_degree of 'O' -- will not result in Jail time are removed (only two of them).
    - We filtered the underlying data from Broward county to include only those rows representing people who had either recidivated in two years, or had at least two years outside of a correctional facility.
- Get new filed longer length of stay
- Get the summary of race, gender, age, xtabs by sex and race
- Plot the data with race and decile score

Task 3: Predict racial Bias

- Change some variables(age, race, gender) into factors, and run a logistic regression, comparing low scores to high scores.

Task 4: Read and prepare the data for Risk of Violent Recidivism

- Read the non violent dataset to read the number of rows
- Remove rows based on following conditions
    - If the charge date of a defendants Compas scored crime was not within 30 days from when the person was arrested, we assume that because of data quality reasons, that we do not have the right offense.
    - We coded the recidivist flag -- is_recid -- to be -1 if we could not find a compas case at all.
    - In a similar vein, ordinary traffic offenses -- those with a c_charge_degree of 'O' -- will not result in Jail time are removed (only two of them).
    - We filtered the underlying data from Broward county to include only those rows representing people who had either recidivated in two years, or had at least two years outside of a correctional facility.
- Get new filed longer length of stay
- Get the summary of race, age category
- Plot the data with race and decile score

Task 5: Predict accuracy

- In order to test whether Compas scores do an accurate job of deciding whether an offender is Low, Medium or High risk, We used the counting model and removed people when they were incarcerated. Due to errors in the underlying jail data, we need to filter out 32 rows that have an end date more than the start date. Considering that there are 13,334 total rows in the data, such a small amount of errors will not affect the results
- Read the cox-parsed.csv dataset to read the number of rows
- Get summary of score factor and race factor
- Test algorithm for Logistic regression on Black and white defendants
- Get summary of fit, white fit and black fit
- Get summary of coxph for white and black data

Task 6: Directions of racial bias

- Read cox-parsed.csv
- Print white and black descendants

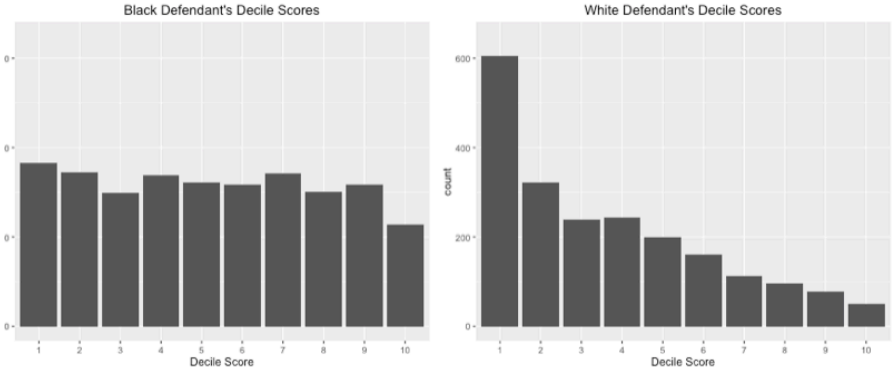
Task 7: Risk of Violent recidivism

- Read cox-violent parsed.csv
- Print white and black descendants

6. Project Evidences

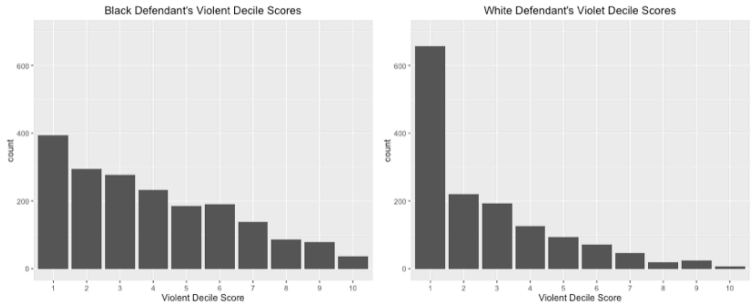
The Learner has to submit the following evidences

A Project Report which comprises of the screen shots of each and every activity to show that the Tasks of each activity has been executed correctly.

Evidence checklist	Summary of expected evidence required by Learner
Task 1	IRAC Objectives
Task 2	Summary of Sex, xtabs, race Black defendants: 51.44% White defendants: 34.07% Hispanic defendants: 8.25% Asian defendants: 0.50% Native American defendants: 0.18% Female 1175 Male 4997 race sex African-American Asian Caucasian Hispanic Native American Other Female 549 2 482 82 2 58 Male 2626 29 1621 427 9 285 Plot decile scores 
Task 3	Racial bias screenshot Call: glm(formula = score_factor ~ gender_factor + age_factor + race_factor + priors_count + crime_factor + two_year_recid, family = "binomial", data = df) Deviance Residuals: Min 1Q Median 3Q Max -2.9966 -0.7919 -0.3303 0.8121 2.6024 Coefficients: Estimate Std. Error z value Pr(>|z|) (Intercept) -1.52554 0.07851 -19.430 < 2e-16 *** gender_factorFemale 0.22127 0.07951 2.783 0.005388 ** age_factorGreater than 45 -1.35563 0.09908 -13.682 < 2e-16 *** age_factorLess than 25 1.30839 0.07593 17.232 < 2e-16 *** race_factorAfrican-American 0.47721 0.06935 6.881 5.93e-12 *** race_factorAsian -0.25441 0.47821 -0.532 0.594717 race_factorHispanic -0.42839 0.12813 -3.344 0.000827 *** race_factorNative American 1.39421 0.76612 1.820 0.068784 . race_factorOther -0.82635 0.16208 -5.098 3.43e-07 *** priors_count 0.26895 0.01110 24.221 < 2e-16 *** crime_factorM -0.31124 0.06655 -4.677 2.91e-06 *** two_year_recid 0.68586 0.06402 10.713 < 2e-16 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 (Dispersion parameter for binomial family taken to be 1) Null deviance: 8483.3 on 6171 degrees of freedom Residual deviance: 6168.4 on 6160 degrees of freedom
Task 4	Summary of race , age category

```
African-American            Asian        Caucasian         Hispanic
            1918               26             1459              355
 Native American            Other
               7              255

25 - 45 Greater than 45    Less than 25
   2300              954             766
```

Plot violent decile scores

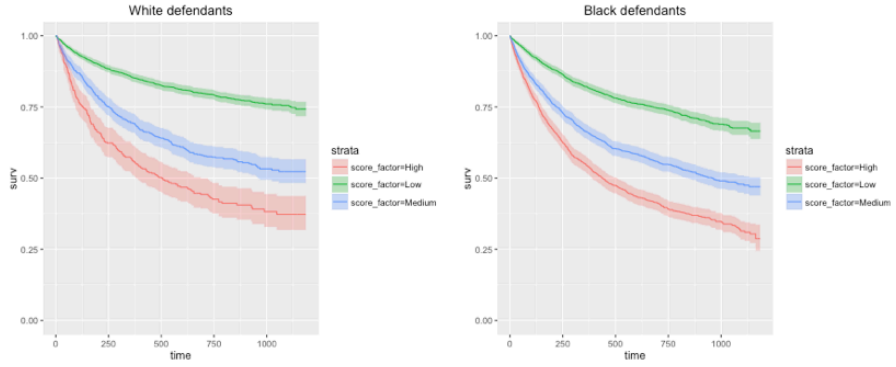


Task 5

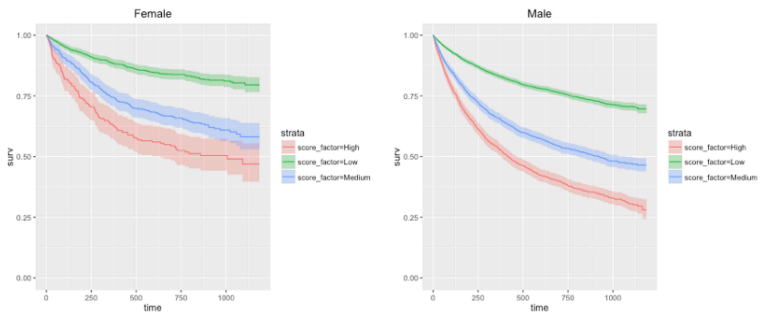
Summary of coxph

```
Call:
coxph(formula = f, data = data)

  n= 13344, number of events= 3469

                    coef exp(coef) se(coef)     z Pr(>|z|)
score_factorHigh   1.24969   3.48927  0.04146 30.14   <2e-16 ***
score_factorMedium 0.79627   2.21725  0.04077 19.53   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                   exp(coef) exp(-coef) lower .95 upper .95
score_factorHigh       3.489     0.2866     3.217     3.785
score_factorMedium     2.217     0.4510     2.047     2.402

Concordance= 0.636  (se = 0.004 )
Rsquare= 0.068   (max possible= 0.99 )
Likelihood ratio test= 942.8  on 2 df,   p=0
Wald test            = 954.8  on 2 df,   p=0
Score (logrank) test = 1055  on 2 df,   p=0
```

Task 6



Task 7



	Black defendants Low High Survived 1692 1043 0.86 Recidivated 170 273 0.14 Total: 3178.00 False positive rate: 38.14 False negative rate: 38.37 Specificity: 0.62 Sensitivity: 0.62 Prevalence: 0.14 PPV: 0.21 NPV: 0.91 LR+: 1.62 LR-: 0.62 White defendants Low High Survived 1679 380 0.91 Recidivated 129 77 0.09 Total: 2265.00 False positive rate: 18.46 False negative rate: 62.62 Specificity: 0.82 Sensitivity: 0.37 Prevalence: 0.09 PPV: 0.17 NPV: 0.93 LR+: 2.03 LR-: 0.77

7. Project Guidelines

You should follow the below guidelines while implementing the Project:

- Implement the project in the technical environment specified in the Project brief
- Follow the format specified for Project report
- The project report should be submitted at least 2 days before the date of Summative Assessment date
- Present the Milestones in every Mentoring Session and seek the Mentor's feedback and review. Incorporate the feedback in your project.
- Attach all project evidences for each milestone as part of your Project report
- During the summative assessment, you will present your project using Project report and where required you will demonstrate the process and evidences through the Project Report.

8. Project Technical Environment

The Learner should perform the project using Jupyter notebook, Python, Azure Machine Learning Studio and the specified dataset.

9. Structure of Project Report

Prepare a Project report with the following index and contents:

- Requirement Analysis
- Why IRAC is a suitable solution for this scenario
- How to: Import and Prepare Data into Jupyter notebook (Loading data)
- Racial bias
- Risk of Violent Recidivism
- Predictive Accuracy
- Directions of the Racial Bias

- IRAC Process : Explain the process of IRAC
- Setting up the Data Preprocessing: Explain the Process of loading, Summary of different fields and plot the data based on race (2 – 3 pages)
- Code for Data loading racial bias : Attach the Code used in the Project for Data Loading
- Code for risk of violent recidivism
- Code for predict accuracy and explain the process

– Code for directions of racial bias and explain the process

You should explain how you have performed each of the above tasks (at least one page per task) and the modules used in each of the above activity. Use Screen captures of the experiment, where required