Data Description Report

Responsible AI, Team 6

Roberto Yulee – Project Manager, Subject Matter Expert

Stephanie DeMaria – Statistician, Data Analyst

Maeve McCarty – Data Engineer

Quanxin Zhang – Data Scientist

Xiaocun Zhu – Technical Analyst

Table of Contents

DEFINING AI BIAS

Algorithmic bias occurs when an algorithm produces inaccurate outcomes as a result of systemic prejudice due to erroneous presumptions in the machine learning process. In other words, when the AI model yields different outcomes or predictions for units that are virtually the same barring sensitive attributes that should hold no correlation. An example of this would be the Apple credit card algorithm. Apple implemented an algorithm that was meant to accurately assess credit to those who applied for a card. Instead it became the most high profile case of AI bias to date. Instead of accurately and fairly assessing credit scores to applicants, the algorithm began to discriminate against female applicants, and lend bias towards men even in cases where women had better credit scores than the men.

An example of bias within our datasets would be in the IBM employee attrition dataset, where there are higher rates of predicted attrition attributed to sensitive variable gender or marital status.

Overall there are three main points of bias. As stated in the AI Fairness 360 codebase they are:

1. Pre-processing - Outcomes in the training data set are biased towards specific instances
2. In-processing - Models are biased towards specific input attributes
3. Post-processing - The test data set is biased towards correct answers that may be biased

DATA SOURCE DESCRIPTION

The following data sets will be used throughout the project to test our bias detection algorithm: IBM HR Employee Attrition, Adult Data, and COMPAS Scores. These three data sets all contain AI prediction model outputs that are known to be biased. Therefore, we will use the data to test and ensure the algorithm we create 1) detects bias correctly and 2) displays the results so that it is easily understood. Within each set, the data will be split into a training or testing group to create and run the algorithm.

A. IBM HR Employee Attrition

The IBM HR Employee Attrition data set will serve as our primary data set. Accenture Federal gave us this csv file to use as a frame of reference to check our work. The model predicted if an employee would attrite (or not) depending on their marital status or gender.

Important Variables

- Attrition (Yes/No) - The predictor outcome that is required to test if bias exists
- Marital Status (Single/Married/Divorced) - One of the sensitive attributes we're measuring the possible bias of
- Gender (Male/Female) - One of the sensitive attributes we're measuring the possible bias of

Data Table: [Data Tables on Github]
- 1470 x 35
- Data Dictionary

B. Adult Data

The Adult Data will serve as our secondary testing data. Accenture Federal also gave us this data set to experiment with, so we know the results are biased. It was found from UC Irvine's Machine Learning Repository and was extracted by Barry Becker from the 1994 Census database. The model that was used predicted whether one's income would exceed $50,000 per year based on race and gender.

Acquisition: Converted a data file into a csv file using Excel.

Important Variables
- Income (>50K/<=50K) - The predictor outcome that is used and required to test if bias exists
- Race (White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black) - One of the sensitive attributes we're measuring the possible bias of
- Gender (Male/Female) - One of the sensitive attributes we're measuring the possible bias of

Data Table: Data Tables on Github
- 32561 x 15
- Data Dictionary

C. COMPAS Scores

Ideally, this data set will be used as a final test to see if our bias detection algorithm works. This data set (csv file) was found on Github. It contains information used by the risk assessment software known as COMPAS to predict which criminals are most likely to reoffend. It is often cited as an example of bias in AI, has open source analyses to refer to/check with, and specifically deals with a relevant sensitive attribute: race.

Important Fields
- Decile Score (1-10) - The quantitative predictor outcome that is used and required to test if bias exists
- Score Text (Low/Medium/High) - The qualitative predictor outcome that is used and required to test if bias exists
- Race (African American, Asian, Caucasian, Hispanic, Native American) - The sensitive attribute we're measuring the possible bias of

Data Table: Data Table on Github
- COMPAS-scores-raw.csv

- 60844 x 28
- [Data Dictionary]

DATA MANIPULATION AND MUNGING

A. IBM Employee HR Attrition

Quality

The IBM HR Employee Attrition Data is very high quality. After further research, we found that it is a fictional dataset created by IBM scientists. Therefore, the data is both organized very well and clearly defined in the data dictionary.

Our dataset contains a total of 1470 rows. It has 35 columns (categorical and quantitative), and the column *Attrition* is used as the target column. This is identified as a BinaryClassification problem.

Reformatting

We are exploring the SHAP, LIME, and AIF360 packages. These packages and almost all types of analysis require numeric variables. To achieve this, we converted nine of the original thirty-five features from categorical variables to quantitative variables. We used Scikit-learn label encoding to encode the character data.

```
Feature:  Attrition
{'No': 0, 'Yes': 1}
Feature:  BusinessTravel
{'Non-Travel': 0, 'Travel_Frequently': 1, 'Travel_Rarely': 2}
Feature:  Department
{'Human Resources': 0, 'Research & Development': 1, 'Sales': 2}
Feature:  EducationField
{'Human Resources': 0, 'Life Sciences': 1, 'Marketing': 2, 'Medical': 3, 'Other': 4, 'Technical Degree': 5}
Feature:  Gender
{'Female': 0, 'Male': 1}
Feature:  JobRole
{'Healthcare Representative': 0, 'Human Resources': 1, 'Laboratory Technician': 2, 'Manager': 3,
'Manufacturing Director': 4, 'Research Director': 5, 'Research Scientist': 6, 'Sales Executive': 7, 'Sales
Representative': 8}
Feature:  MaritalStatus
{'Divorced': 0, 'Married': 1, 'Single': 2}
Feature:  Over18
{'Y': 0}
Feature:  OverTime
{'No': 0, 'Yes': 1}
```

Cleaning/Missing Data

There were no outliers or missing values.

Variables

Several new variables were created for the IBM employee attrition dataset. The new variables include: GenderMaritalStatus, GenderJobRole, GenderJobSatisfaction, GenderYearsAtCompany, GenderYearsInCurrentRole, GenderEducationField, GenderTrainingTimeLastYear, GenderHourlyRateLevel, HourlyRateLevel, and MaritalStatusHourlyRateLevel.

The five variables 'DailyRate', 'EmployeeCount', 'EmployeeNumber', 'MonthlyRate', and 'Over18' were dropped due to redundancy or failure to add significant information. For instance, the 'Over 18' was dropped since every instance was a yes.

No Additional Data Sources

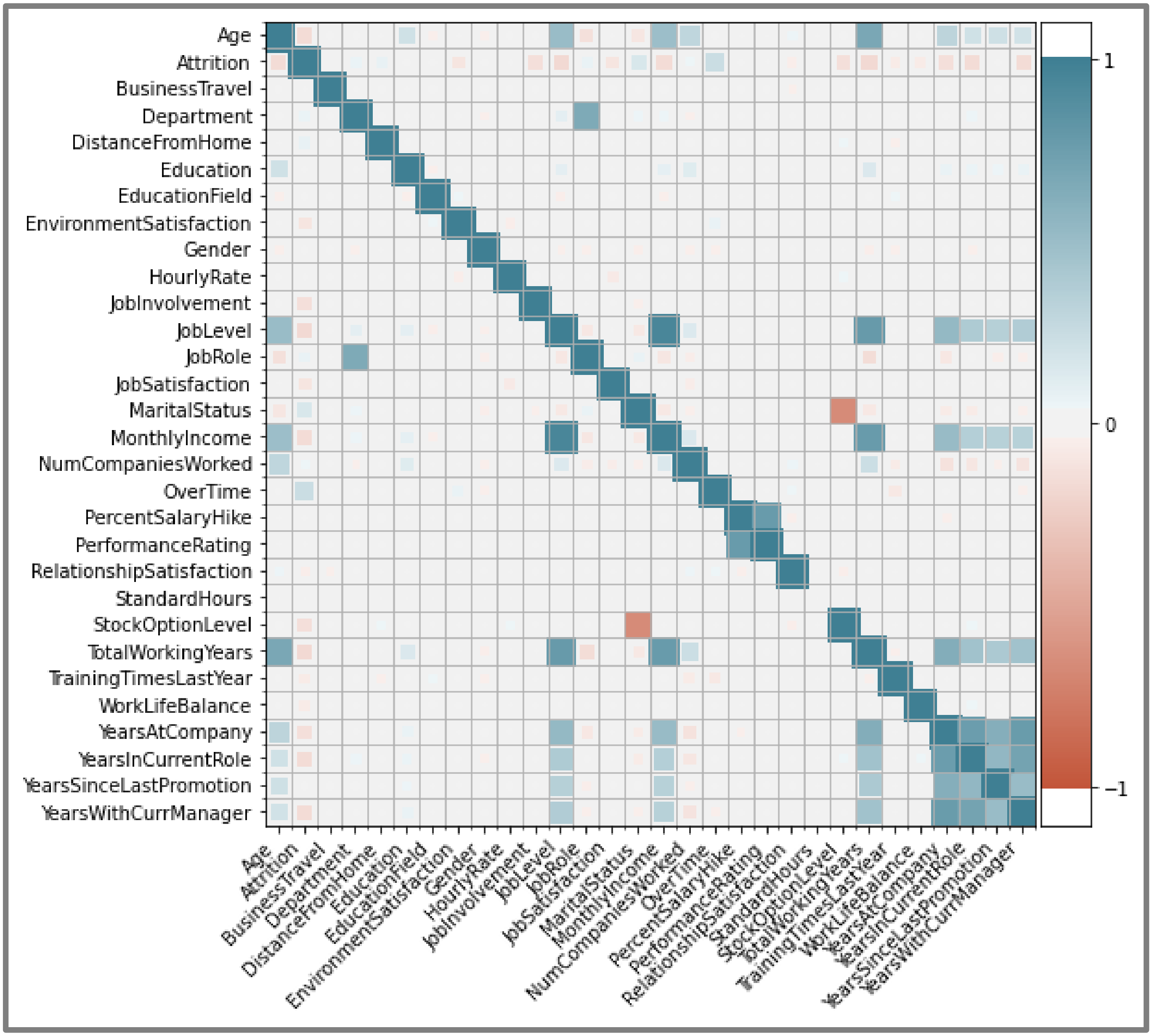
This data set was created by IBM scientists, so little munging efforts were required.

Data Exploration Effort

Descriptive Statistics

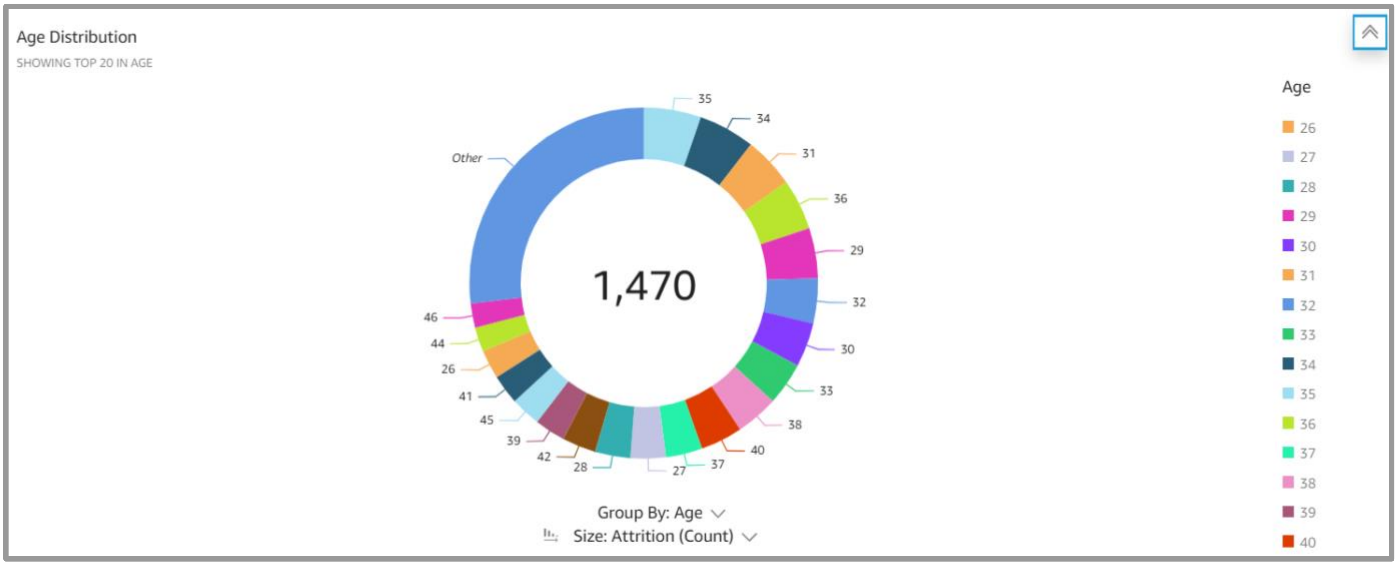
	count	mean	std	min	25%	50%	75%	max
Age	1470.0	36.923810	9.135373	18.0	30.0	36.0	43.00	60.0
Attrition	1470.0	0.161224	0.367863	0.0	0.0	0.0	0.00	1.0
BusinessTravel	1470.0	1.607483	0.665455	0.0	1.0	2.0	2.00	2.0
Department	1470.0	1.260544	0.527792	0.0	1.0	1.0	2.00	2.0
DistanceFromHome	1470.0	9.192517	8.106864	1.0	2.0	7.0	14.00	29.0
Education	1470.0	2.912925	1.024165	1.0	2.0	3.0	4.00	5.0
EducationField	1470.0	2.247619	1.331369	0.0	1.0	2.0	3.00	5.0
EnvironmentSatisfaction	1470.0	2.721769	1.093082	1.0	2.0	3.0	4.00	4.0
Gender	1470.0	0.600000	0.490065	0.0	0.0	1.0	1.00	1.0
HourlyRate	1470.0	65.891156	20.329428	30.0	48.0	66.0	83.75	100.0
JobInvolvement	1470.0	2.729932	0.711561	1.0	2.0	3.0	3.00	4.0
JobLevel	1470.0	2.063946	1.106940	1.0	1.0	2.0	3.00	5.0
JobRole	1470.0	4.458503	2.461821	0.0	2.0	5.0	7.00	8.0
JobSatisfaction	1470.0	2.728571	1.102846	1.0	2.0	3.0	4.00	4.0
MaritalStatus	1470.0	1.097279	0.730121	0.0	1.0	1.0	2.00	2.0
MonthlyIncome	1470.0	6502.931293	4707.956783	1009.0	2911.0	4919.0	8379.00	19999.0
NumCompaniesWorked	1470.0	2.693197	2.498009	0.0	1.0	2.0	4.00	9.0
OverTime	1470.0	0.282993	0.450606	0.0	0.0	0.0	1.00	1.0
PercentSalaryHike	1470.0	15.209524	3.659938	11.0	12.0	14.0	18.00	25.0
PerformanceRating	1470.0	3.153741	0.360824	3.0	3.0	3.0	3.00	4.0
RelationshipSatisfaction	1470.0	2.712245	1.081209	1.0	2.0	3.0	4.00	4.0
StandardHours	1470.0	80.000000	0.000000	80.0	80.0	80.0	80.00	80.0
StockOptionLevel	1470.0	0.793878	0.852077	0.0	0.0	1.0	1.00	3.0
TotalWorkingYears	1470.0	11.279592	7.780782	0.0	6.0	10.0	15.00	40.0
TrainingTimesLastYear	1470.0	2.799320	1.289271	0.0	2.0	3.0	3.00	6.0
WorkLifeBalance	1470.0	2.761224	0.706476	1.0	2.0	3.0	3.00	4.0
YearsAtCompany	1470.0	7.008163	6.126525	0.0	3.0	5.0	9.00	40.0
YearsInCurrentRole	1470.0	4.229252	3.623137	0.0	2.0	3.0	7.00	18.0
YearsSinceLastPromotion	1470.0	2.187755	3.222430	0.0	0.0	1.0	3.00	15.0
YearsWithCurrManager	1470.0	4.123129	3.568136	0.0	2.0	3.0	7.00	17.0

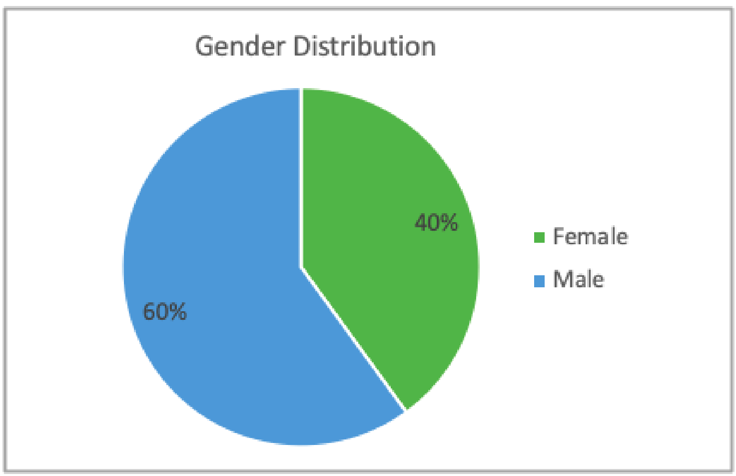
Correlation Analysis

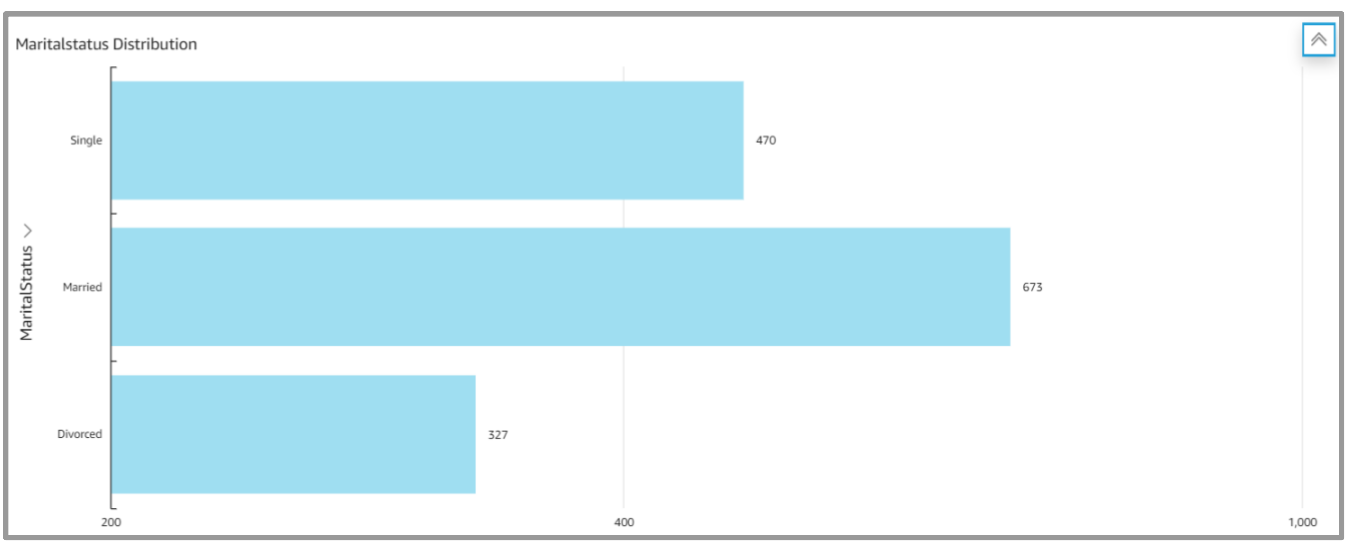


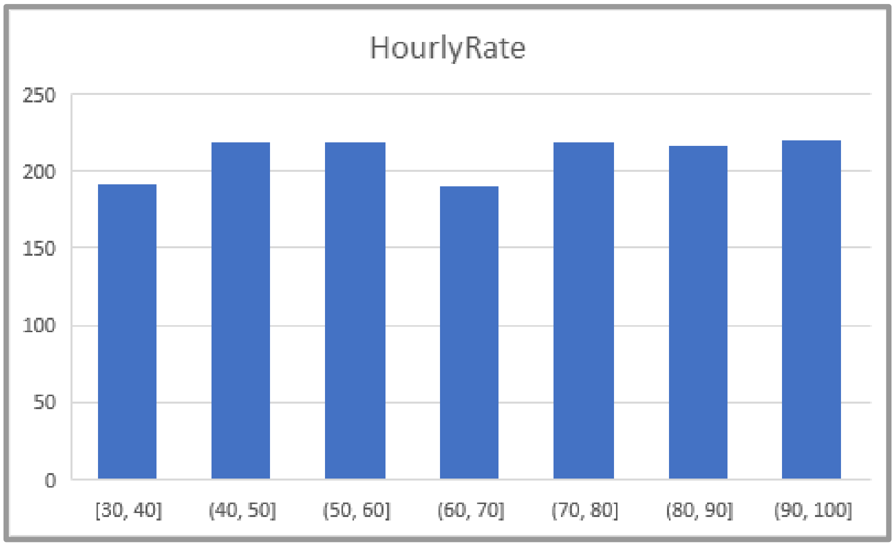
Key Variables to Examine:

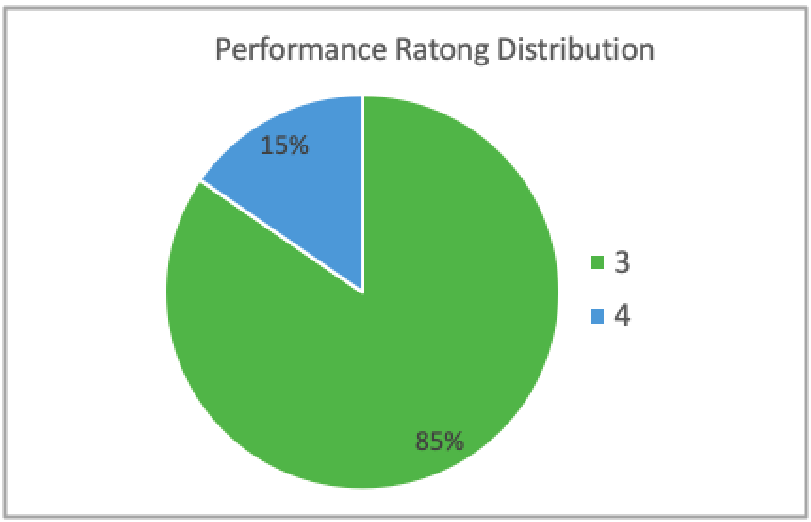
1. Age: Did those who leave tend to be older or younger?
2. Gender: Is gender a bias factor?
3. Marital Status: Do personal relationships and family affect attrition?
4. Hourly Rate: Is hourly rate a key factor that makes employees leave?
5. Performance: Did those who leave the job score lower on performance?
6. Education Field: Which education field is more likely to leave?

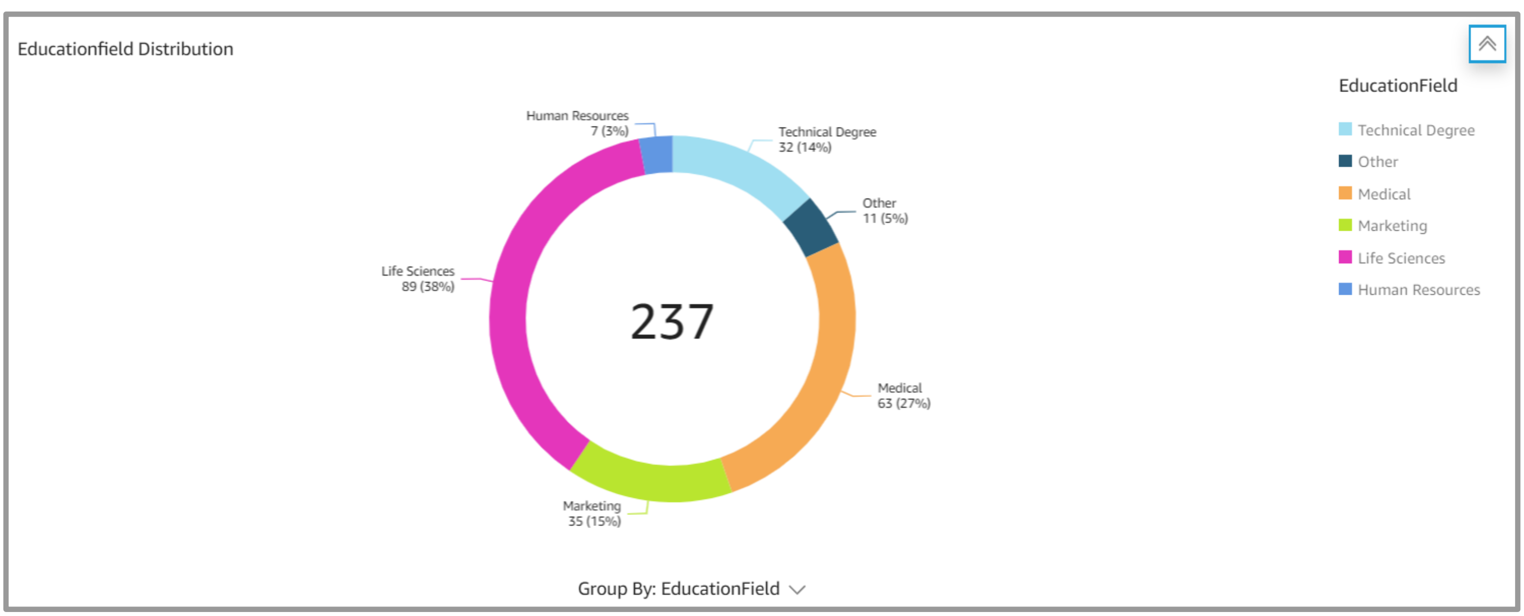

Age Distribution
SHOWING TOP 20 IN AGE
1,470
Group By: Age
Size: Attrition (Count)
Age
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40

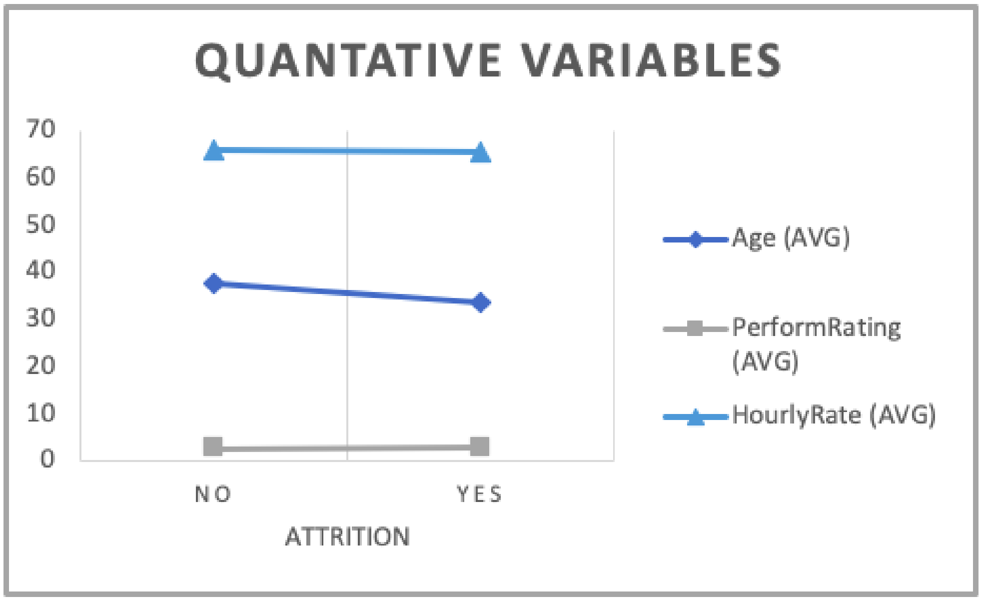

Gender Distribution
40%
60%
Female
Male

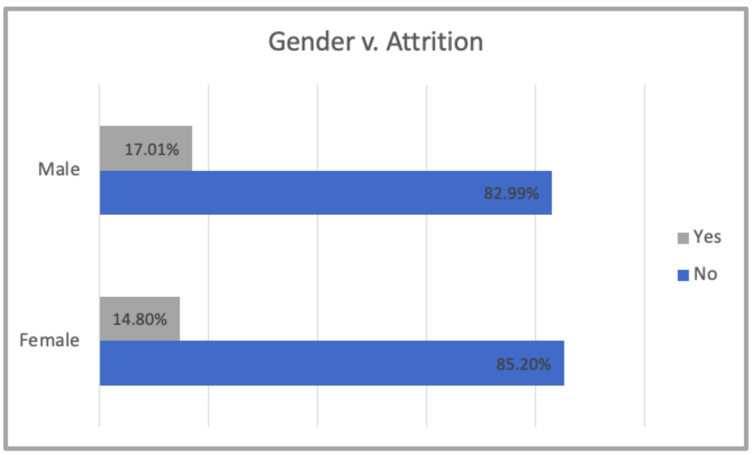

Maritalstatus Distribution
MaritalStatus
Single
Married
Divorced
470
673
327
200
400
1,000

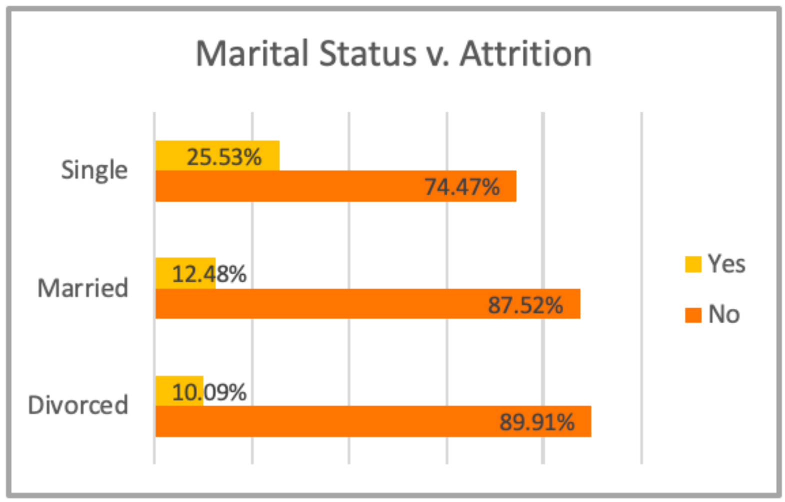

HourlyRate
250
200
150
100
50
0
[30, 40]
(40, 50]
(50, 60]
(60, 70]
(70, 80]
(80, 90]
(90, 100]

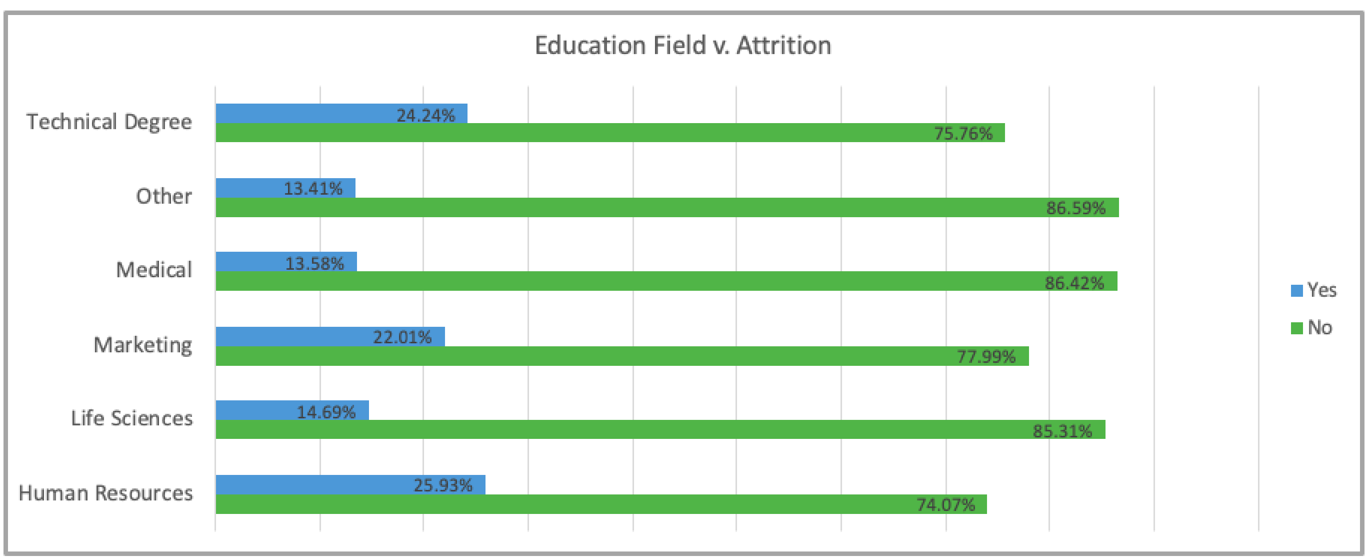

Performance Ratong Distribution
15%
85%
3
4

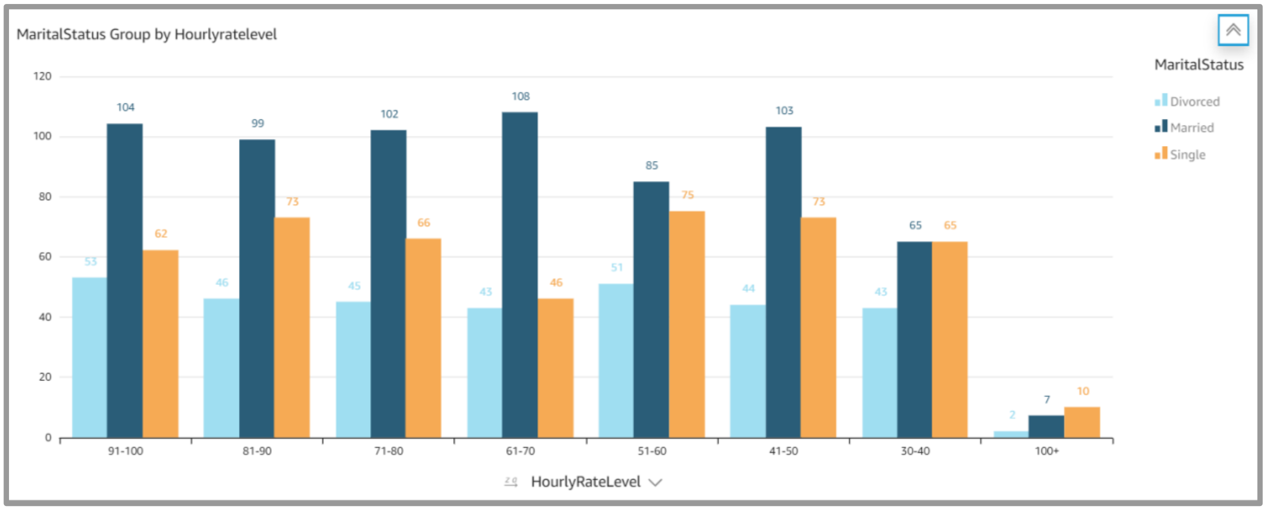

Educationfield Distribution
Human Resources
7 (3%)
Technical Degree
32 (14%)
Other
11 (5%)
Life Sciences
89 (38%)
237
Medical
63 (27%)
Marketing
35 (15%)
EducationField
Technical Degree
Other
Medical
Marketing
Life Sciences
Human Resources
Group By: EducationField

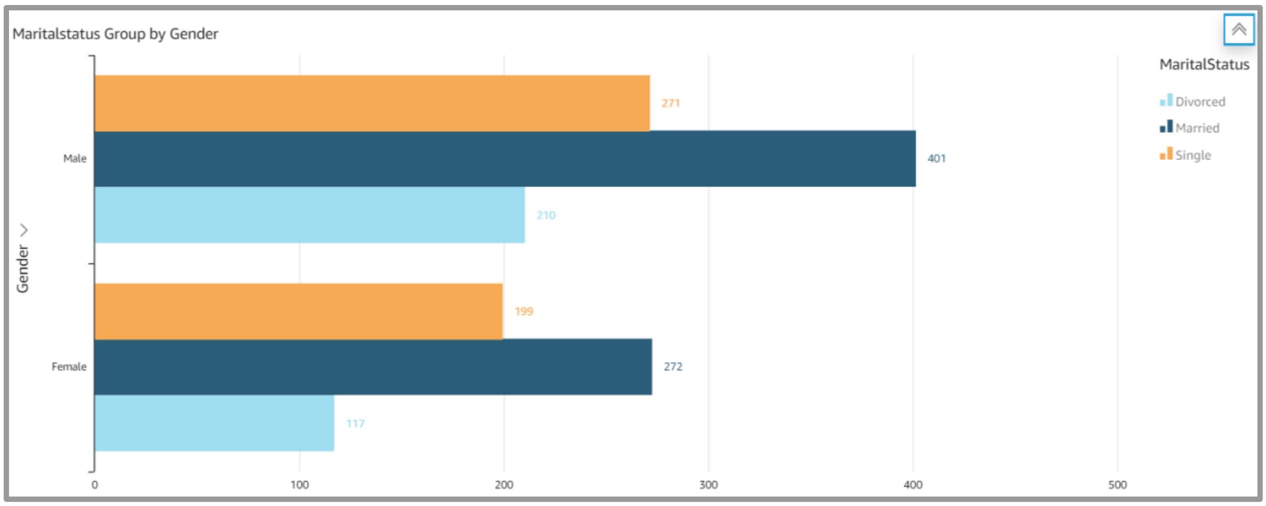

QUANTATIVE VARIABLES
70
60
50
40
30
20
10
0
NO
YES
ATTRITION
Age (AVG)
PerformRating (AVG)
HourlyRate (AVG)


Gender v. Attrition
Male
17.01%
82.99%
Female
14.80%
85.20%
Yes
No


Marital Status v. Attrition
Single
25.53%
74.47%
Married
12.48%
87.52%
Divorced
10.09%
89.91%
Yes
No


Education Field v. Attrition
Technical Degree
24.24%
75.76%
Other
13.41%
86.59%
Medical
13.58%
86.42%
Marketing
22.01%
77.99%
Life Sciences
14.69%
85.31%
Human Resources
25.93%
74.07%
Yes
No





B. Adult Data

Quality

The Adult Data is above average quality. The data is both organized very well and clearly defined in the data dictionary.

The dataset contains a total of 32561 rows. It has 9 columns (categorical and quantitative), and the column *predictedIncome (<=50K or >50K)* is used as the target column. This is identified as a BinaryClassification problem.

Missing Data

The adult data set has 4,262 missing values which resulted in about 2% of the overall dataset being missing. Since the percentage of missing values is less than 5%, it is recommended to drop the data with missing values. The rows with missing values were dropped in excel by using the find and replace feature. First, the find function in excel identified all the cells with missing values by searching for the placeholder value which was “?”. After all of the missing values were selected in the spreadsheet, the row to which they belonged could be easily dropped. Our new CSV file contains no missing data and is ready to be used for data munging.

Another alternative is to use python. Using functions .isnull().sum() and .dropna, we were able to find and delete the missing values from the data set.

Reformatting

We are exploring the SHAP, LIME, and AIF360 packages. These packages and almost all types of analysis require numeric variables. To achieve this, we converted nine of the original fifteen features from categorical variables to quantitative variables. We used Scikit-learn label encoding to encode the character data.

```
Feature:  workclass
{' Federal-gov': 0, ' Local-gov': 1, ' Private': 2, ' Self-emp-inc': 3, ' Self-emp-not-inc': 4, ' State-gov': 5, ' Without-pay': 6}
Feature:  education
{' 10th': 0, ' 11th': 1, ' 12th': 2, ' 1st-4th': 3, ' 5th-6th': 4, ' 7th-8th': 5, ' 9th': 6, ' Assoc-acdm': 7, ' Assoc-voc': 8, ' Bachelors': 9, ' Doctorate': 10, ' HS-grad': 11, ' Masters': 12, ' Preschool': 13, ' Prof-school': 14, ' Some-college': 15}
Feature:  maritalStatus
{' Divorced': 0, ' Married-AF-spouse': 1, ' Married-civ-spouse': 2, ' Married-spouse-absent': 3, ' Never-married': 4, ' Separated': 5, ' Widowed': 6}
Feature:  occupation
{' Adm-clerical': 0, ' Armed-Forces': 1, ' Craft-repair': 2, ' Exec-managerial': 3, ' Farming-fishing': 4, ' Handlers-cleaners': 5, ' Machine-op-inspct': 6, ' Other-service': 7, ' Priv-house-serv': 8, ' Prof-specialty': 9, ' Protective-serv': 10, ' Sales': 11, ' Tech-support': 12, ' Transport-moving': 13}
Feature:  relationship
{' Husband': 0, ' Not-in-family': 1, ' Other-relative': 2, ' Own-child': 3, ' Unmarried': 4, ' Wife': 5}
Feature:  race
{' Amer-Indian-Eskimo': 0, ' Asian-Pac-Islander': 1, ' Black': 2, ' Other': 3, ' White': 4}
Feature:  sex
{' Female': 0, ' Male': 1}
Feature:  nativeCountry
{' Cambodia': 0, ' Canada': 1, ' China': 2, ' Columbia': 3, ' Cuba': 4, ' Dominican-Republic': 5, ' Ecuador': 6, ' El-Salvador': 7, ' England': 8, ' France': 9, ' Germany': 10, ' Greece': 11, ' Guatemala': 12, ' Haiti': 13, ' Holand-Netherlands': 14, ' Honduras': 15, ' Hong': 16, ' Hungary': 17, ' India': 18, ' Iran': 19, ' Ireland': 20, ' Italy': 21, ' Jamaica': 22, ' Japan': 23, ' Laos': 24, ' Mexico': 25, ' Nicaragua': 26, ' Outlying-US(Guam-USVI-etc)': 27, ' Peru': 28, ' Philippines': 29, ' Poland': 30, ' Portugal': 31, ' Puerto-Rico': 32, ' Scotland': 33, ' South': 34, ' Taiwan': 35, ' Thailand': 36, ' Trinadad&Tobago': 37, ' United-States': 38, ' Vietnam': 39, ' Yugoslavia': 40}
Feature:  predictedIncome
{' <=50K': 0, ' >50K': 1}
```

Variables

No new variables were created or dropped.

No Additional Data Sources
"All data was extracted from the 1994 census database using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0))."

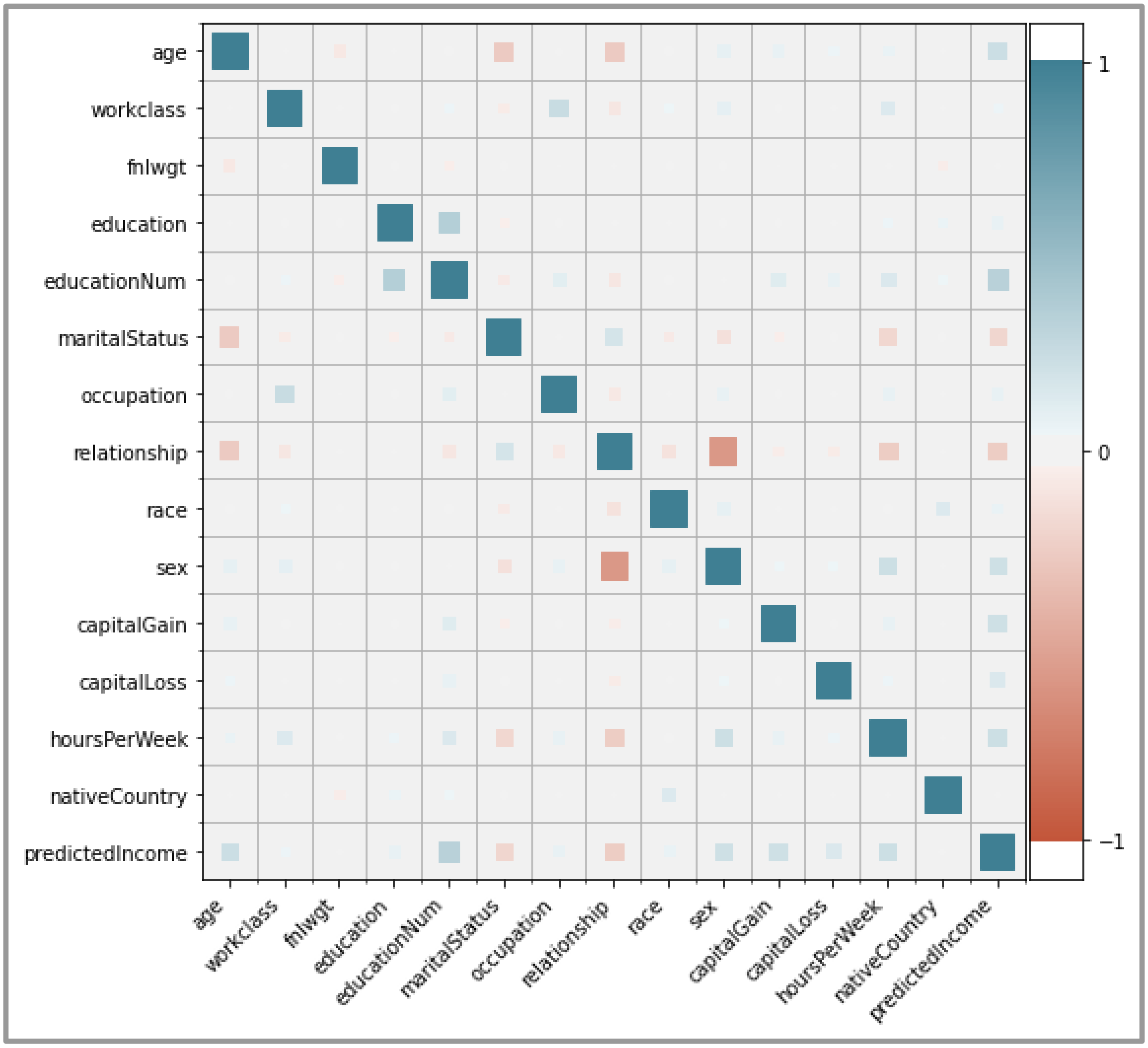
Data Exploration Effort

<u>Descriptive Statistics</u>

	count	mean	std	min	25%
age	30162.0	38.437902	13.134665	17.0	28.00
workclass	30162.0	2.199324	0.953925	0.0	2.00
fnlwgt	30162.0	189793.833930	105652.971529	13769.0	117627.25
education	30162.0	10.333764	3.812292	0.0	9.00
educationNum	30162.0	10.121312	2.549995	1.0	9.00
maritalStatus	30162.0	2.580134	1.498016	0.0	2.00
occupation	30162.0	5.959850	4.029566	0.0	2.00
relationship	30162.0	1.418341	1.601338	0.0	0.00
race	30162.0	3.678602	0.834709	0.0	4.00
sex	30162.0	0.675685	0.468126	0.0	0.00
capitalGain	30162.0	1092.007858	7406.346497	0.0	0.00
capitalLoss	30162.0	88.372489	404.298370	0.0	0.00
hoursPerWeek	30162.0	40.931238	11.979984	1.0	40.00
nativeCountry	30162.0	36.382567	6.105372	0.0	38.00
predictedIncome	30162.0	0.248922	0.432396	0.0	0.00

	50%	75%	max
age	37.0	47.0	90.0
workclass	2.0	2.0	6.0
fnlwgt	178425.0	237628.5	1484705.0
education	11.0	12.0	15.0
educationNum	10.0	13.0	16.0
maritalStatus	2.0	4.0	6.0
occupation	6.0	9.0	13.0
relationship	1.0	3.0	5.0
race	4.0	4.0	4.0
sex	1.0	1.0	1.0
capitalGain	0.0	0.0	99999.0
capitalLoss	0.0	0.0	4356.0
hoursPerWeek	40.0	45.0	99.0
nativeCountry	38.0	38.0	40.0
predictedIncome	0.0	0.0	1.0

Correlation Analysis



Key Variables to Examine:

1. Age: Does age affect the amount of predicted income?
2. Marital Status: Personal relationship and family may affect the predicted income?
3. Race: Is race a prominent factor in determining income?
4. Sex: Is the model biased against male or female when predicting income?

The same steps used to analyze the key variables in the IBM Attrition Data were used to analyze the variables Age, Marital Status, Race and Sex. Graphs have been omitted.

DEVELOPMENT WORKFLOW