Marketing Analytics: Using Global Demographics to Forecast Age and Income

By Tyler Gorecki, Ricky Kuehn,
Doruk Ozar, Luke Schneider, James Siegener

Research Questions

1) **Visualization/Summary EDA**: How does the income distribution look across age groups and education level?

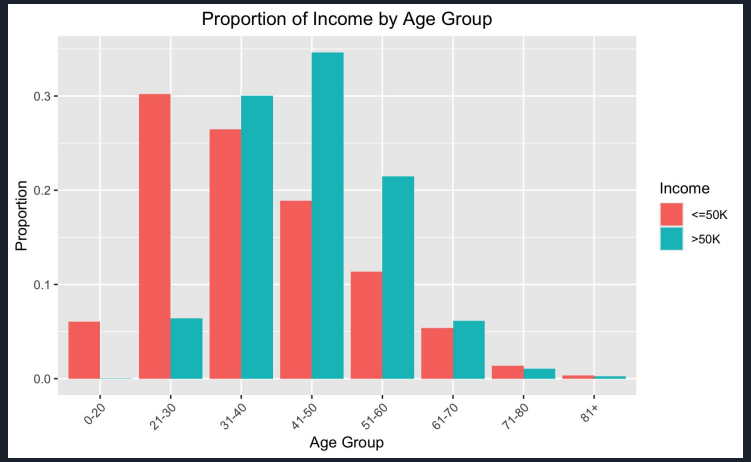
2) **Visualization/Summary EDA**: Are there relationships between predictors of hours worked, education, occupation and income being above 50K?

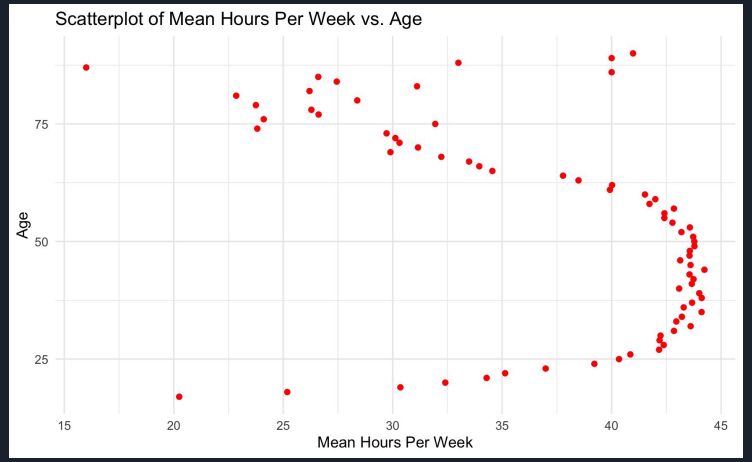
3) **Linear model**: Can age be predicted based on an individual's income level, occupation, education and perhaps other predictors?

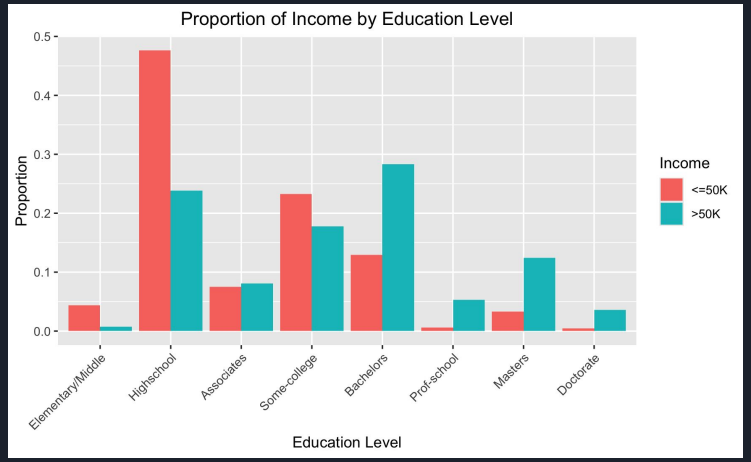
4) **Logistic regression**: Can an individual earning over 50K be predicted based on their age, education level, occupation, hours worked per week, and other available features?

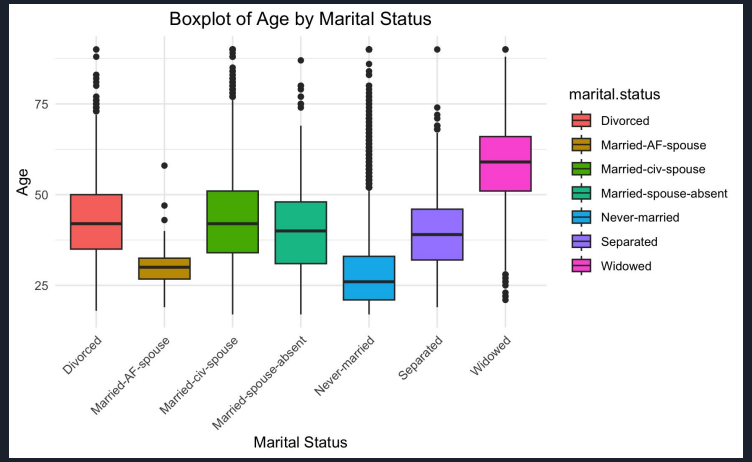
5) **Linear/Log reg use case**: Could age/income predictions be used to isolate individuals ideal for marketing a certain product/service ?

EDA

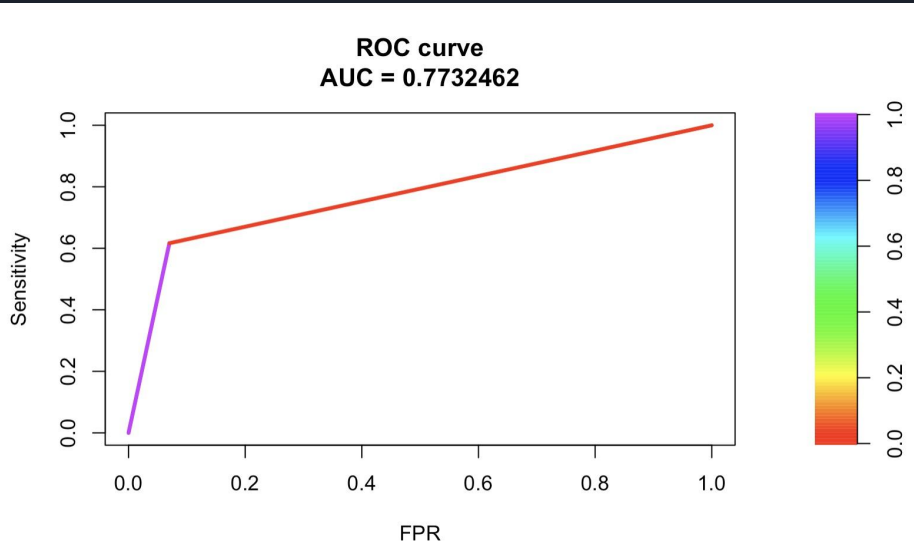

Proportion of Income by Age Group
Proportion
0.0
0.1
0.2
0.3
0-20
21-30
31-40
41-50
51-60
61-70
71-80
81+
Age Group
Income
<=50K
>50K

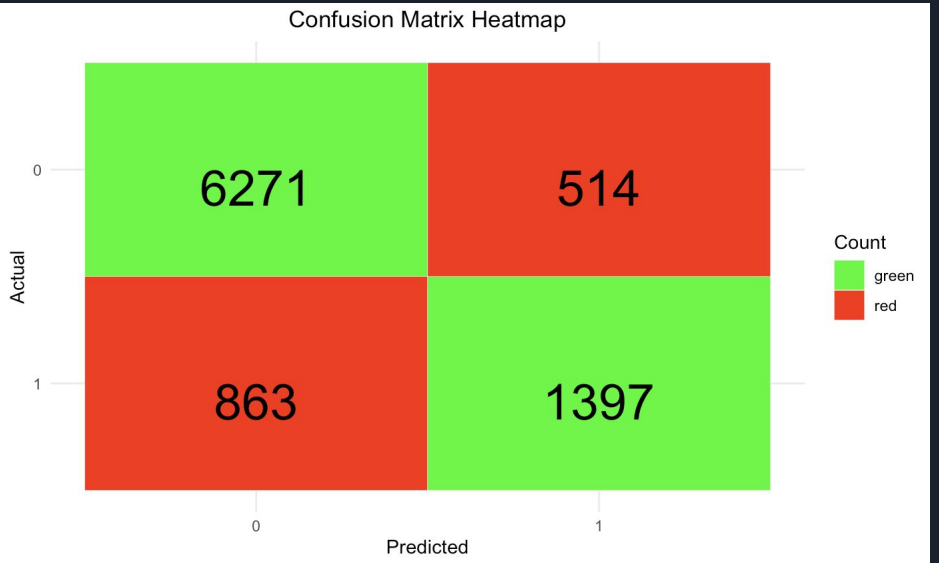

Scatterplot of Mean Hours Per Week vs. Age
Age
25
50
75
15
20
25
30
35
40
45
Mean Hours Per Week


Proportion of Income by Education Level
Proportion
0.0
0.1
0.2
0.3
0.4
0.5
Elementary/Middle
Highschool
Associates
Some-college
Bachelors
Prof-school
Masters
Doctorate
Education Level
Income
<=50K
>50K


Boxplot of Age by Marital Status
Age
25
50
75
Divorced
Married-AF-spouse
Married-civ-spouse
Married-spouse-absent
Never-married
Separated
Widowed
Marital Status
marital.status
Divorced
Married-AF-spouse
Married-civ-spouse
Married-spouse-absent
Never-married
Separated
Widowed

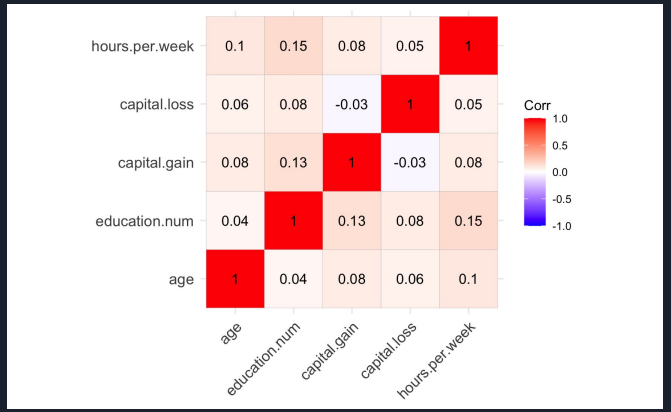
Logistic Regression Model

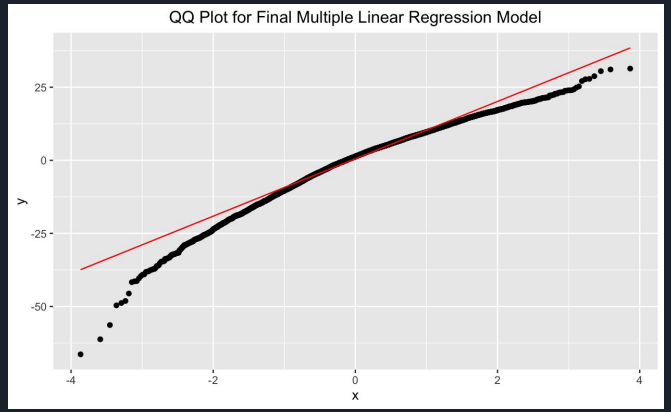


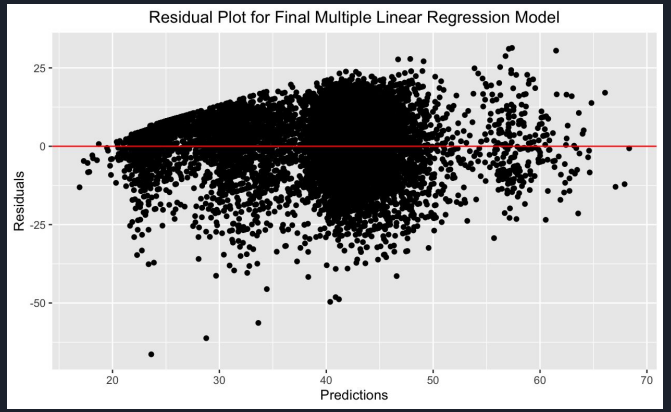


Multiple Linear Regression Model

R^2: 0.39907
RMSE: 10.296

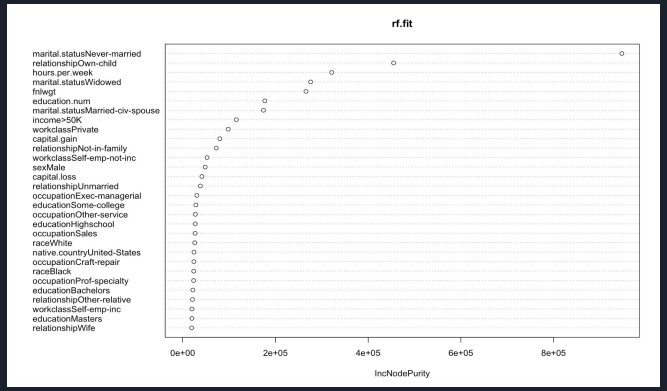


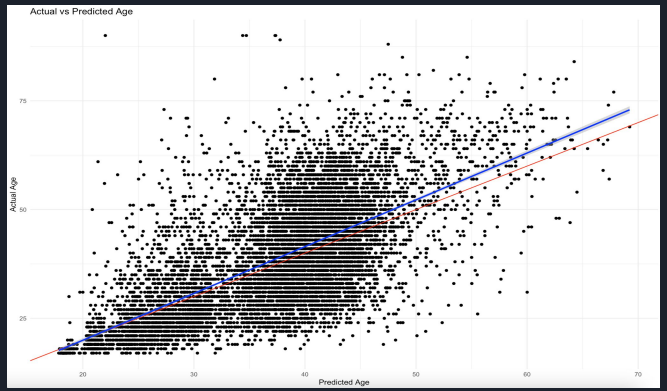




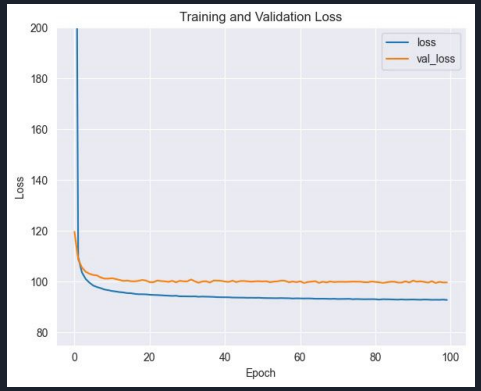
Other Model Attempts

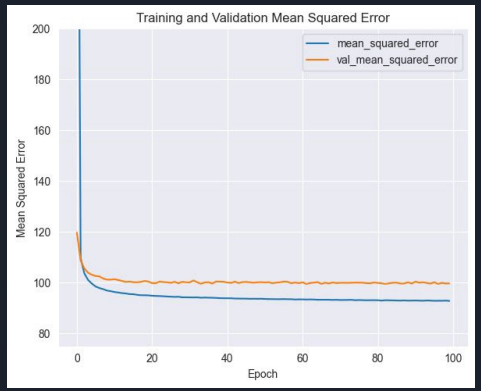
Random Forest





Neural Network



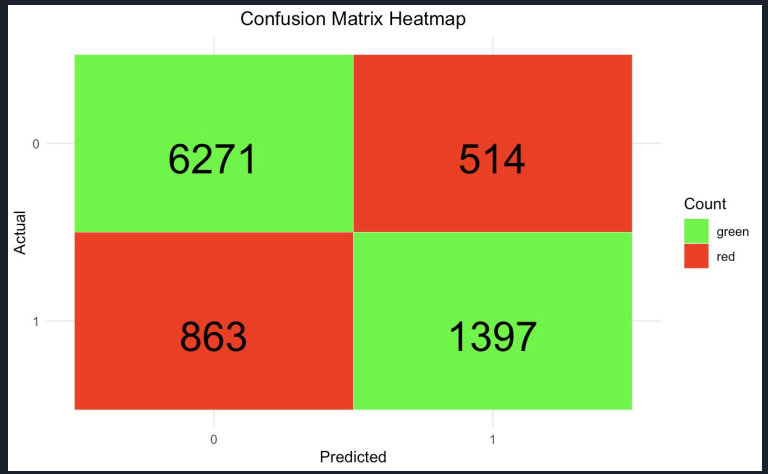


Conclusion

Model	Metric
Logistic Reg.	Accuracy: 85%
Linear Reg.	R^2: 0.39
Random Forest	R^2: 0.51
Neural Network	RMSE: 9.96

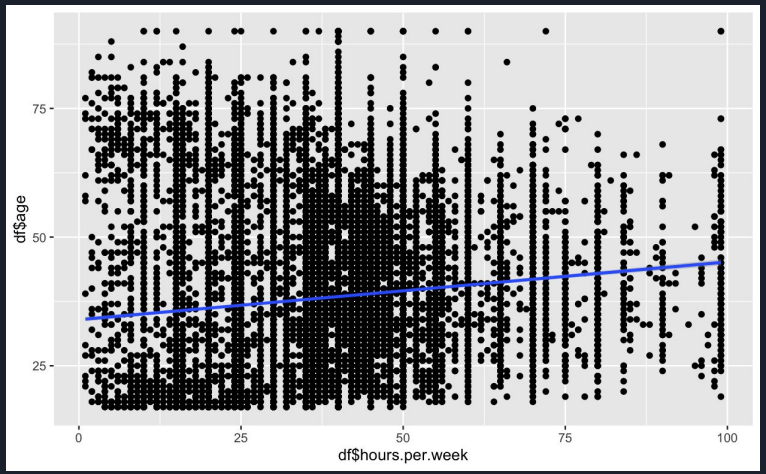
- **Strong binary, poor numerical predictions**
- Future binary prediction outlook:
  - Could change logistic regression thresholds depending on marketing purpose
- Future numeric prediction outlook:
  - Continuous income variable
  - Variable to split age (ex: retirement)

Conclusion



- Strong binary, poor numerical predictions
- **Future binary prediction outlook:**
  - **Could change logistic regression thresholds depending on marketing purpose**
- Future numeric prediction outlook:
  - Continuous income variable
  - Variable to split age (ex: retirement)

Conclusion



- Strong binary, poor numerical predictions
- Future binary prediction outlook:
  - Could change logistic regression thresholds depending on marketing purpose
- **Future numeric prediction outlook:**
  - **Continuous income variable**
  - **Variable to split age (ex: retirement)**

Thank you!