US Census Income Analysis

Group 9

Yuying Chen, Syed Hassan Raza, Brittany Thum, Xinpei Zhao

Executive Summary

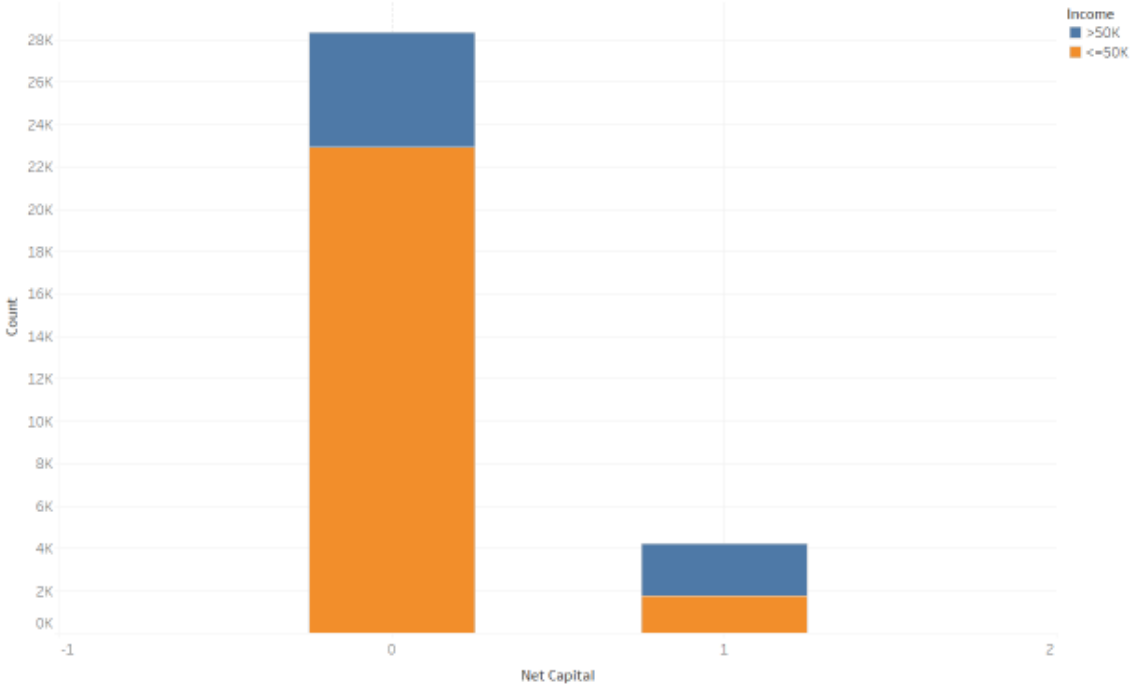
Age, Hours per Week, and Net Capital most significant when predicting income > $50K	5 age segments with 3 main target segments	Ideal Target Customer
Target ages 36-51 who work equal to or greater than 50 hours a week and have high net capital.	Cater campaign channels and messaging towards Adults, Middle Age Adults, and Older Adults	Use ideal target customer characteristics to design campaign

Recommendations

Net Capital (Capital Gain - Capital Loss)

Which variables are the best predictors for Income > $50K?

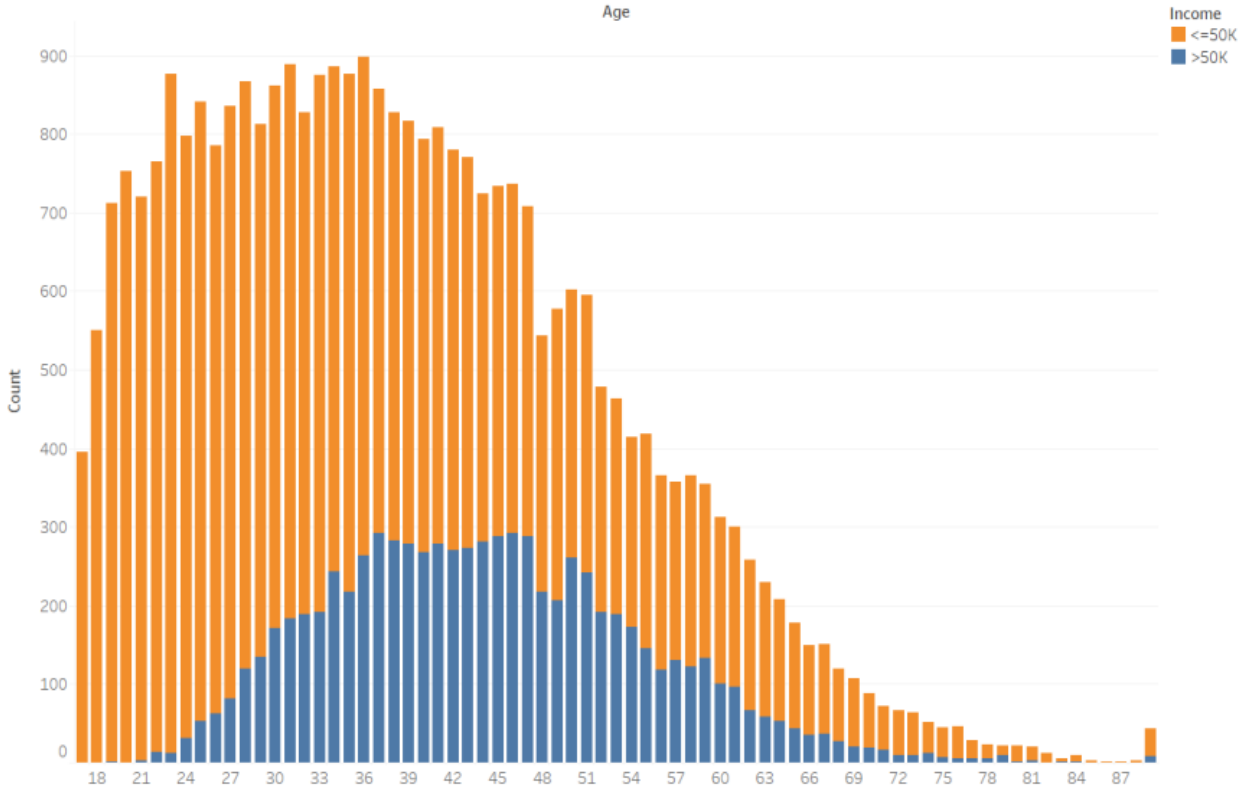
Each unit increase in Net Capital (Capital Gain - Capital Loss) will increase the probability that a household income is greater than 50K by 81.7%



Which variables are the best predictors for Income > $50K?

For each one year increase in Age, there is 50.97% increase in the probability that the income of that person is greater than 50k.

Age and income > $50K peaks between 36-47 and gradually decreases



Which variables are the best predictors for income > $50K?

For each hour per week increase in Hours_per_week work someone does, there is 50.66% increase in the probability that the income of that person is greater than 50k.

*Based on best predictors in each category

What are the 5 Age Segments?

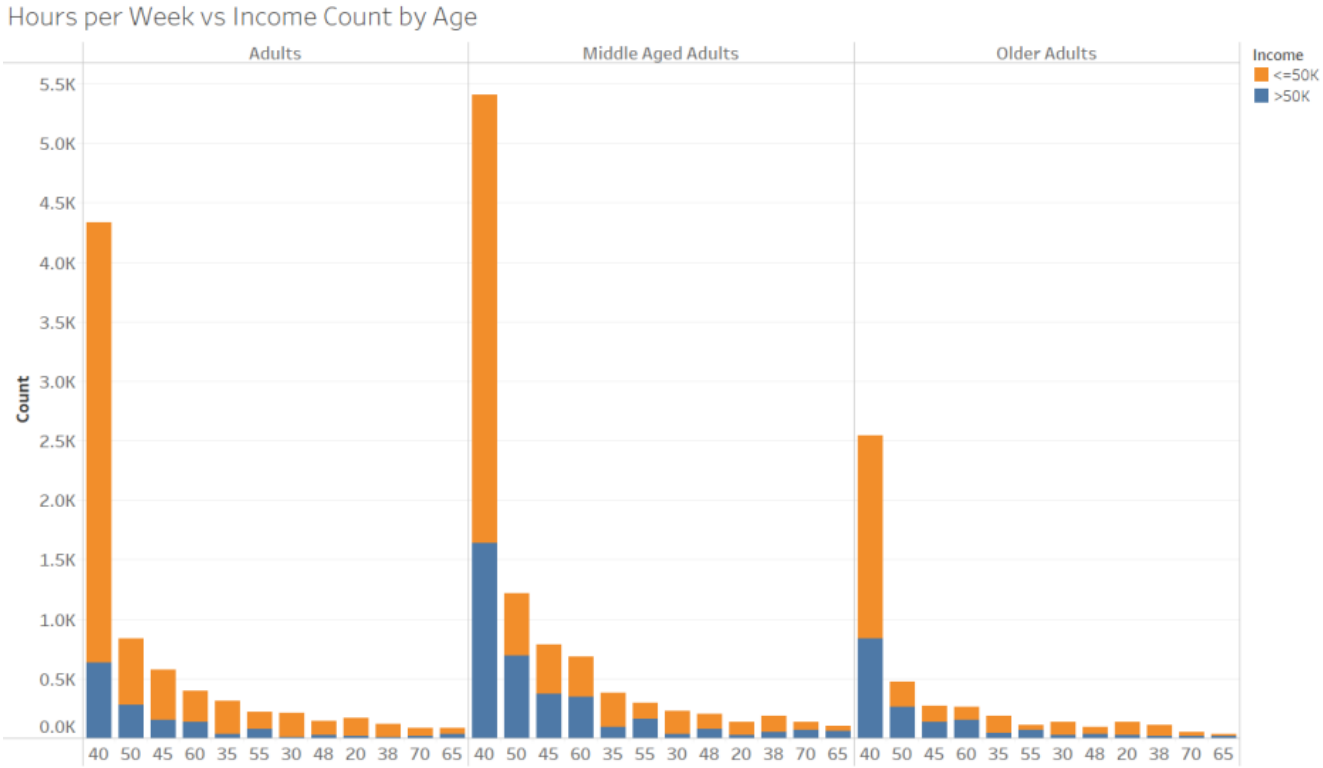
Predictors	Young Adults	Adults	Middle Age Adults	Older Adults	Retirement +
Age	17-25	26-35	36-50	51-64	65+
Education	Bachelor , Some College	Master, Doctorate, Professional School	Master, Doctorate, Professional School	Master, Doctorate, Professional School	Bachelor, Master, Doctorate, Professional School
Relationship	Non-married	Married	Married	Married	Married
Occupation	Professional speciality	Professional speciality	Executive managerial and professional speciality	Executive managerial	Executive managerial
Work Class	Private work class	Private work class	Private work class	Private work class	Private work class
Other Characteristics	College bound, Entry level career	Career focused	Building families	Empty nesters	Retirement

What are the top 3 segments?

Commonalities include

- Higher education
- Private work class
- Married
- Professional Speciality and Executive Management occupations

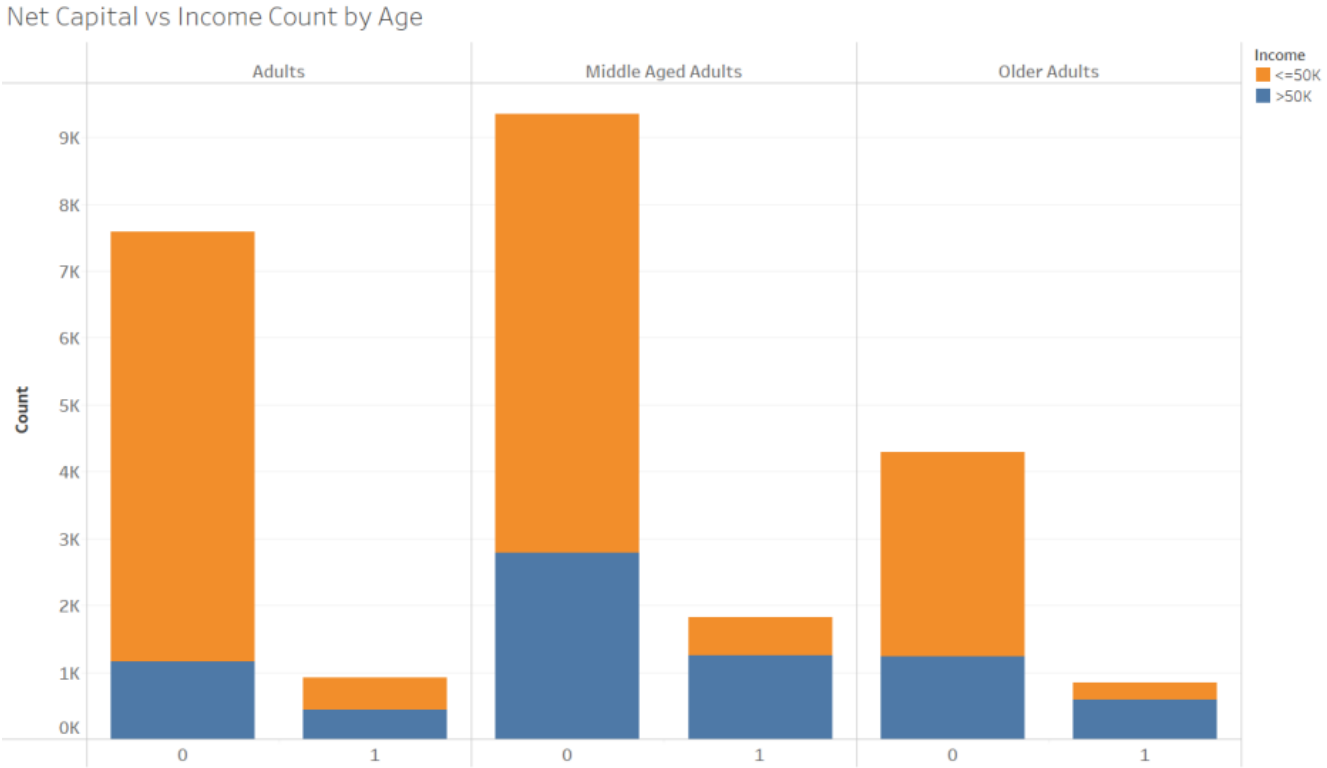
Middle Aged Adults work more hours and earn a higher income.



What are the top 3 segments?

Adults, Middle Age Adults, and Older Adults do not have capital losses and some have a net capital greater than 0.

More Middle Aged Adults have a net capital greater than 0.



Who is the Ideal Customer to target?

Age: 36-51

Sex: Male

Education Level: Masters and Doctorate

Occupation: Executive Managerial, Professional Speciality

Hours per week: Greater than 50 hours

Who is the Ideal Customer to target?

Relationship Status: Married

Race: White, Asian

Native Country/Region: USA and Asia

Net Capital: Higher Capital Gains

Recommendations

1st Recommendation

Target ages 36-51 who work more than 50 hours a week and have high net capital.

2nd Recommendation

Cater campaign channels and messaging towards the 3 age segments.

Different messaging and channels for Adults, Middle Age Adults, and Older Age Adults at different stages of life.

3rd Recommendation

Use ideal target customer characteristics to design campaign and its reach.

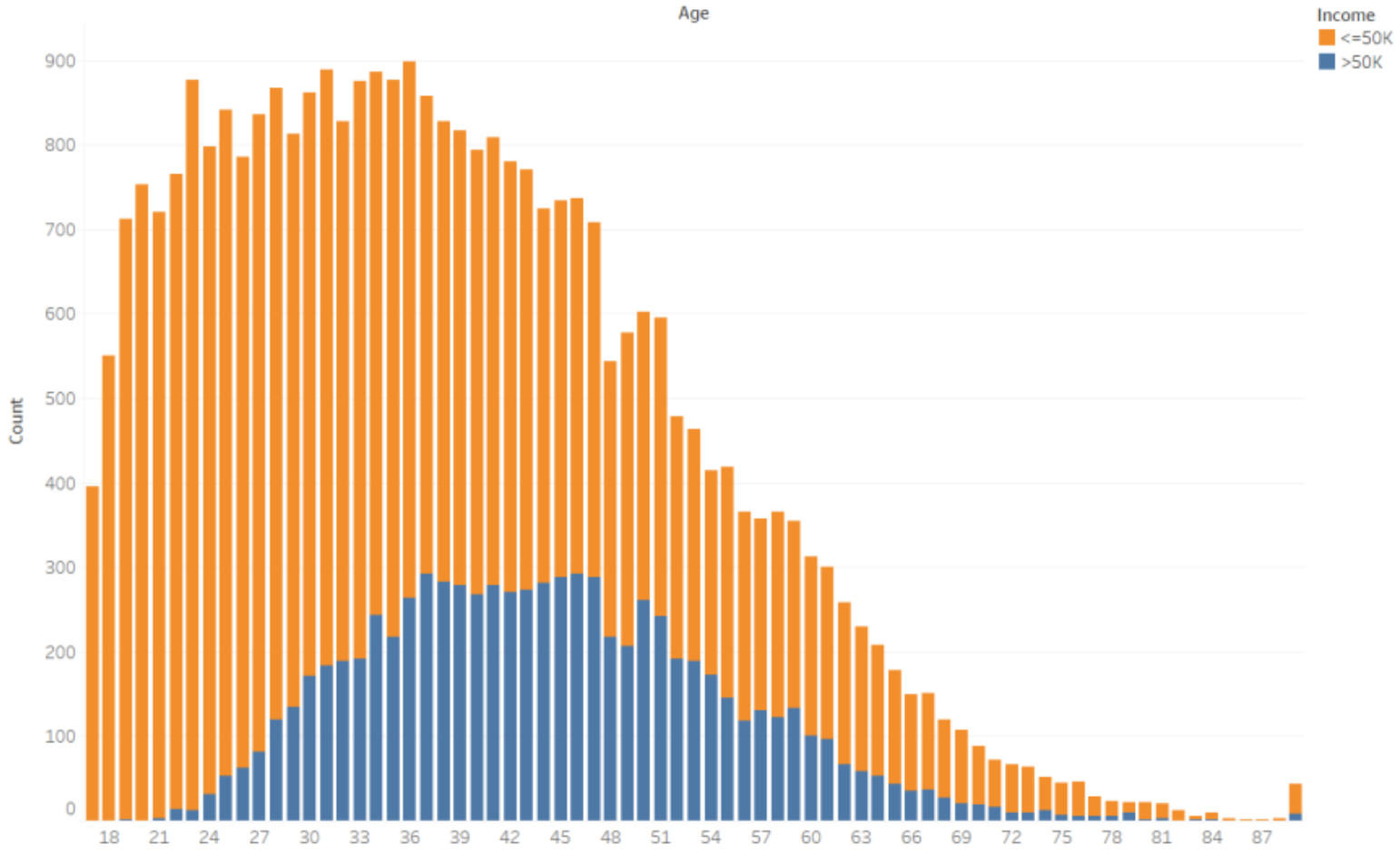
Group 9

Appendix - Model Comparison

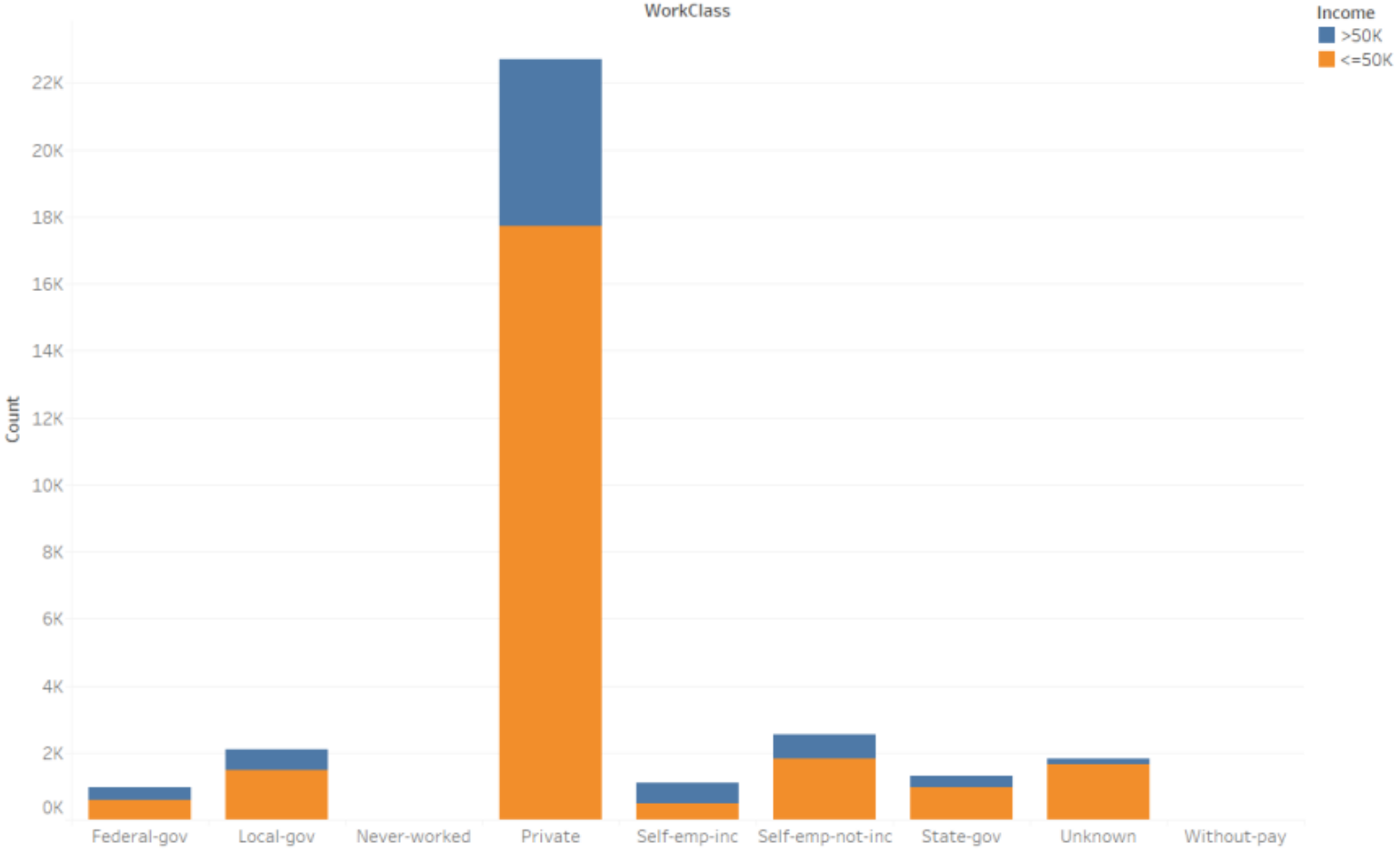
	CV* Score for sample dataset	CV* Score for whole dataset
Naive Bayes	68.13%	67.46%
Logistic Regression	84.40%	84.26%
Decision Tree	96.56%	79.79%
Random Forest	96.56%	81.88%

*Cross validations scores of all models for comparison

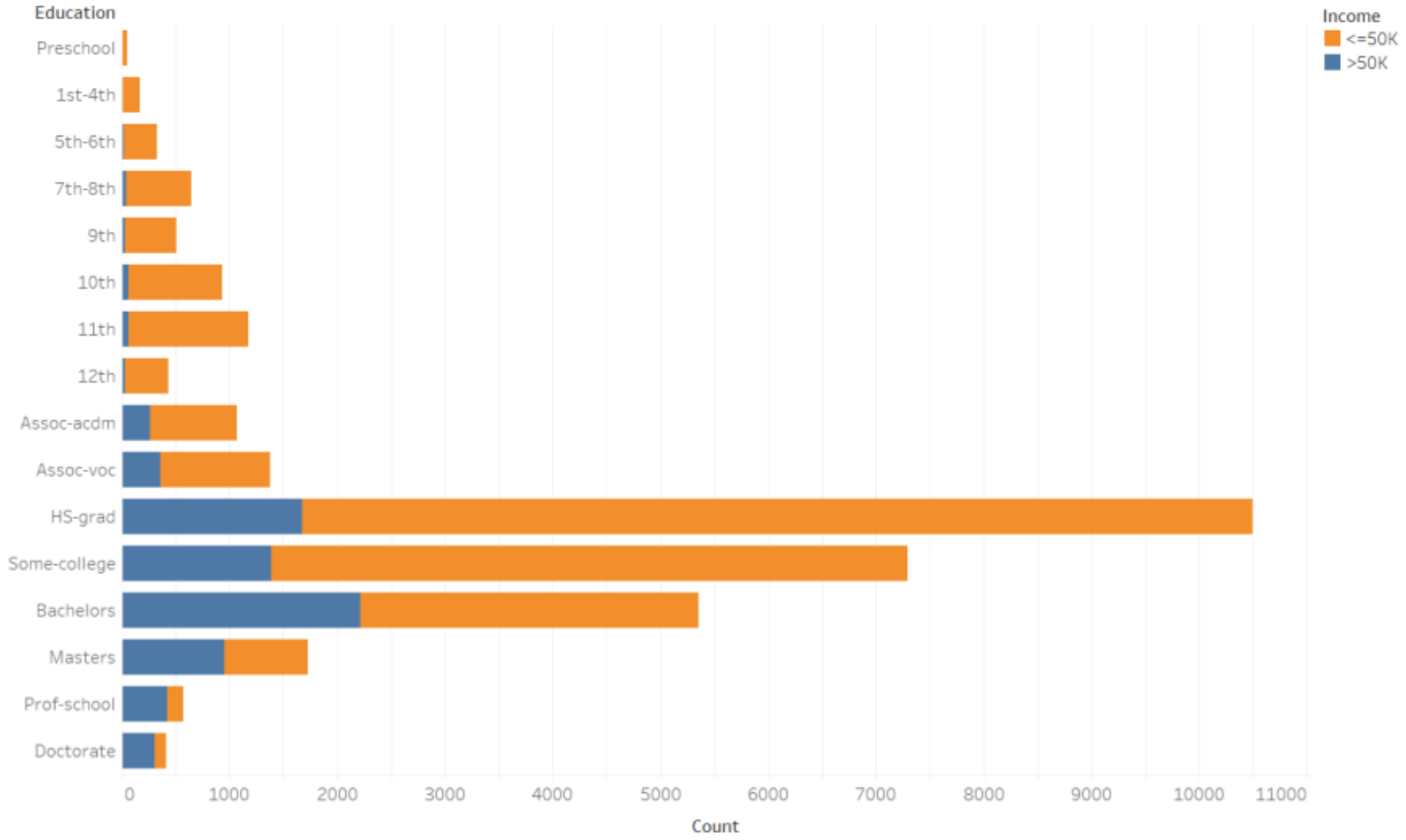
Age v.s. Income



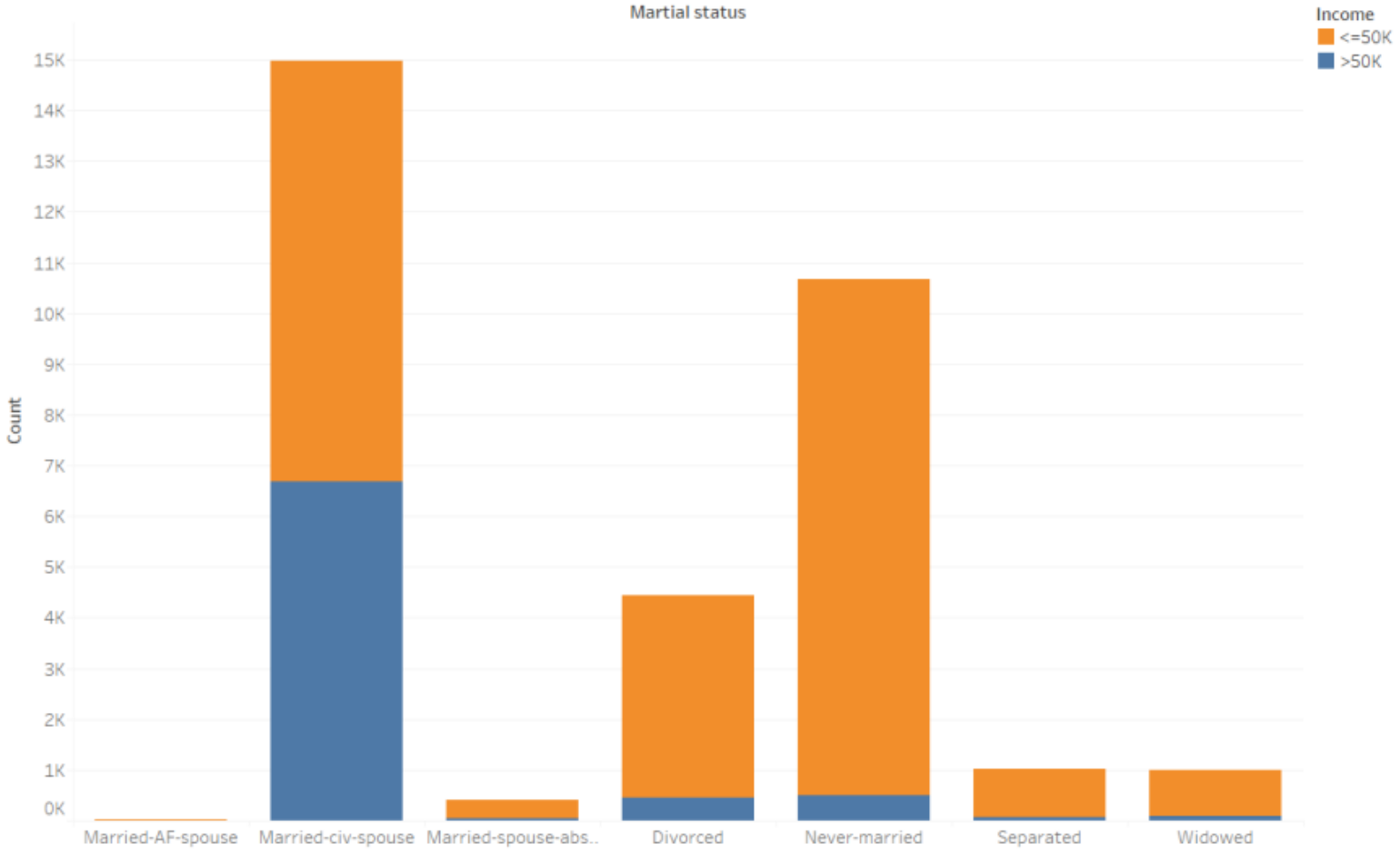
WorkClass v.s. Income



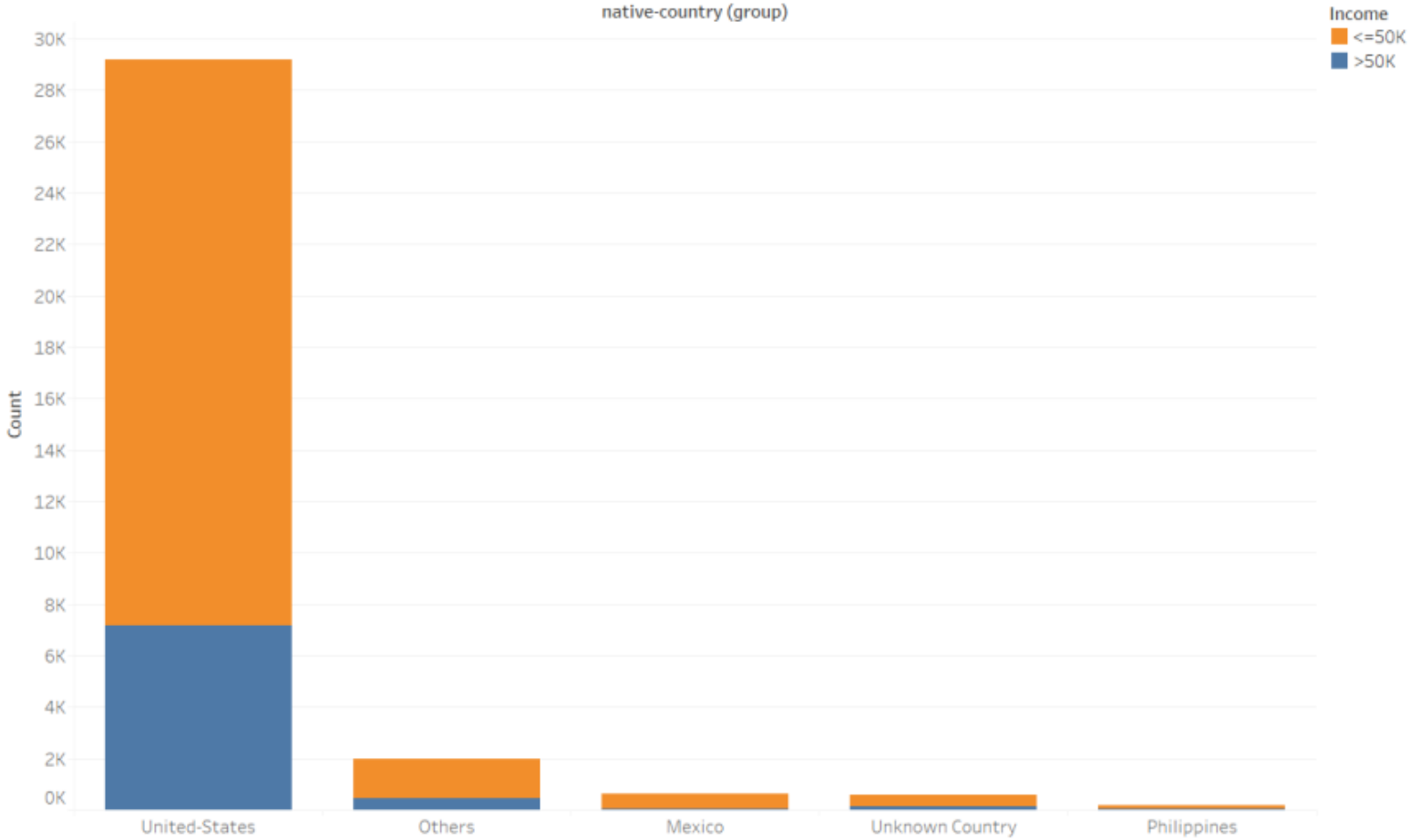
Education Level v.s. Income



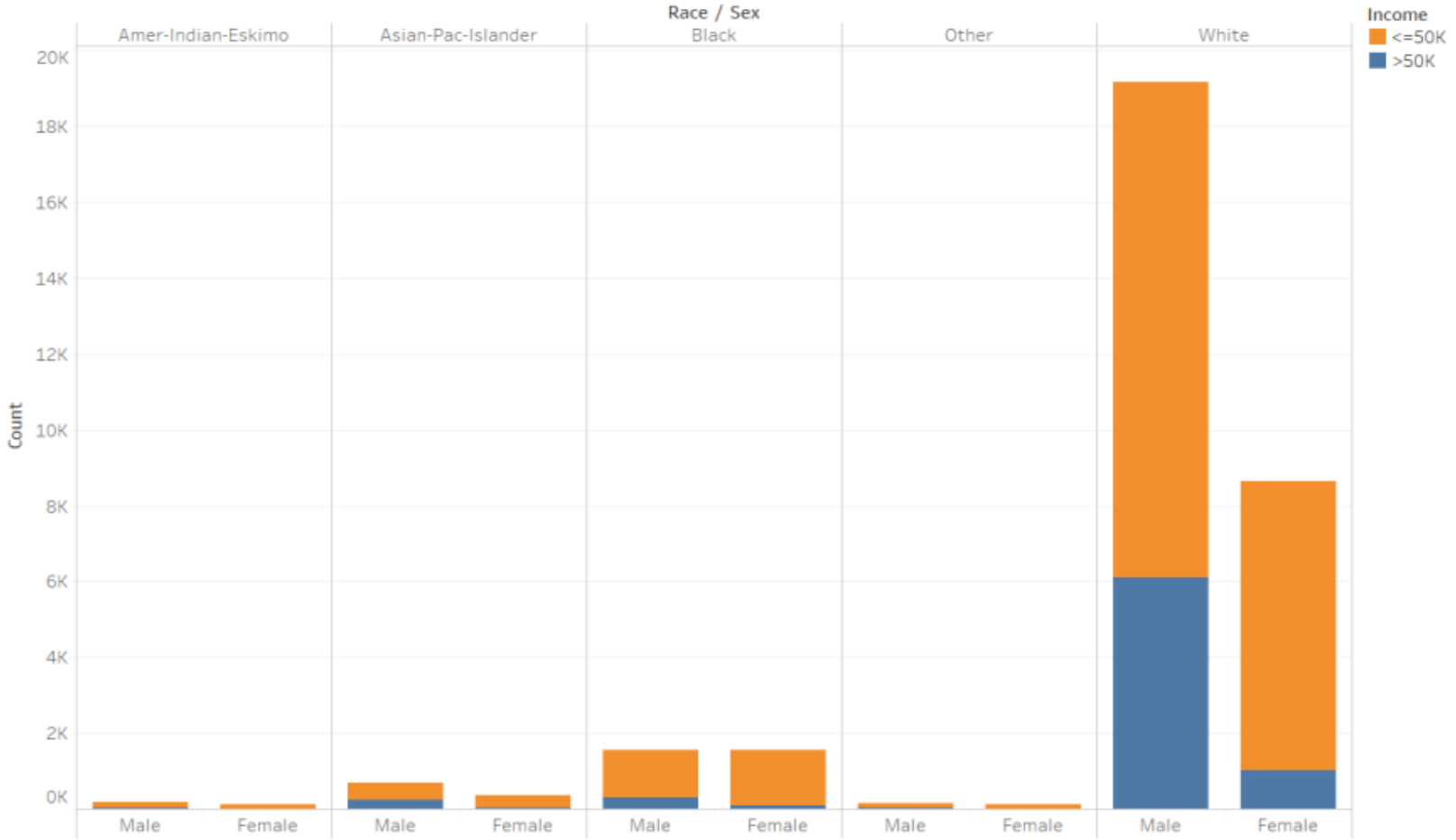
Marital Status v.s. Income



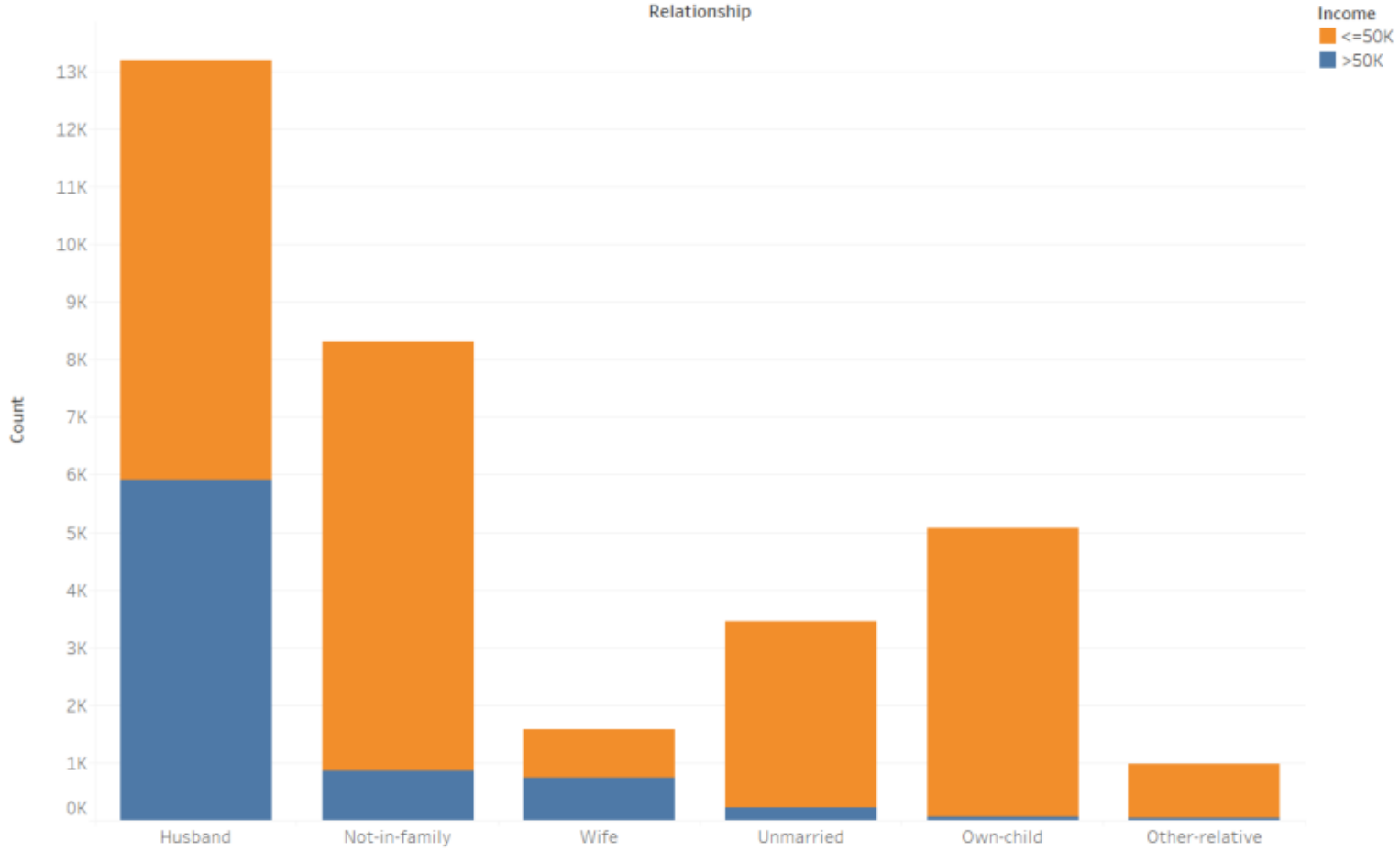
Native Country v.s. Income



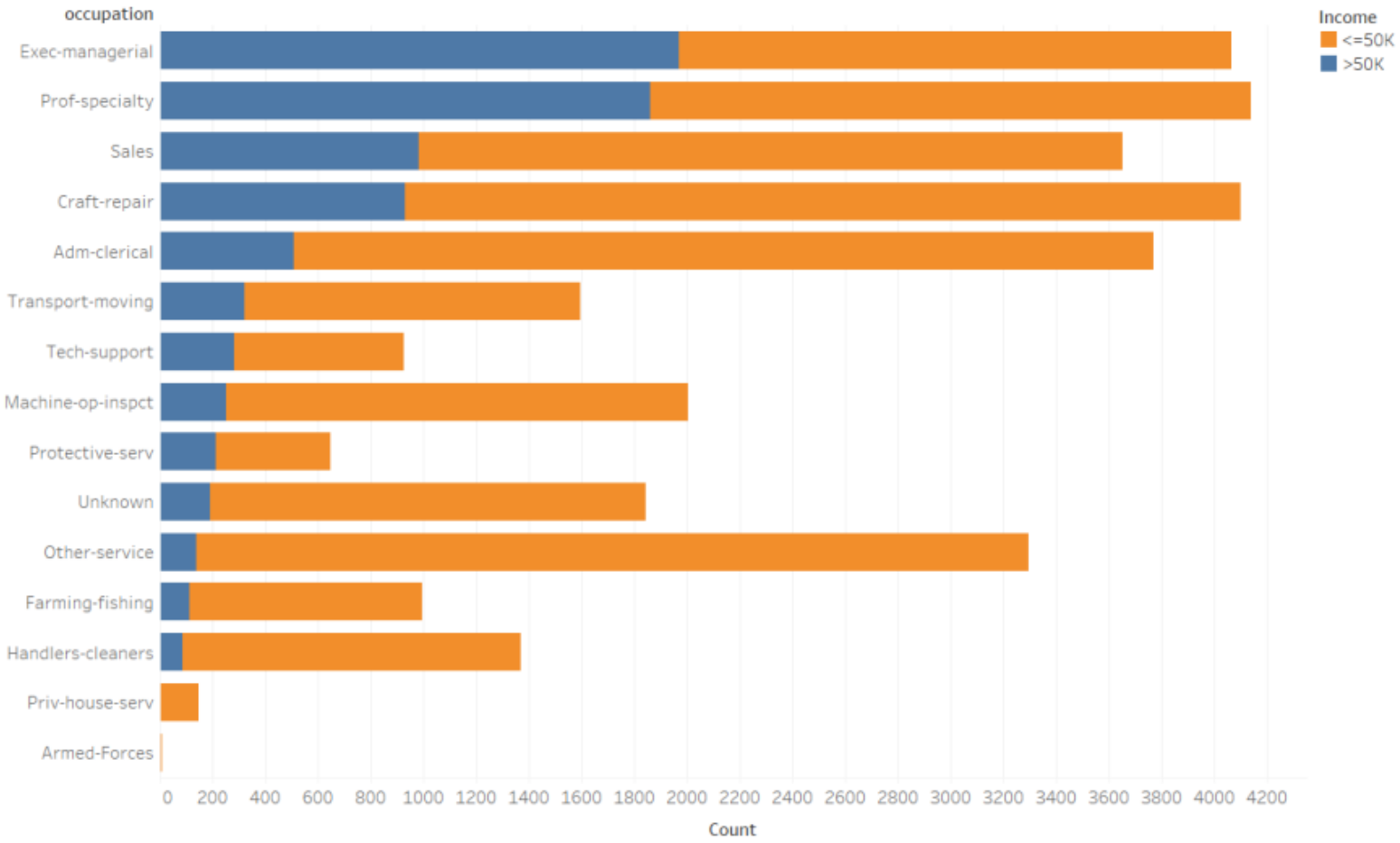
Race/Sex v.s. Income



Relationship v.s. Income

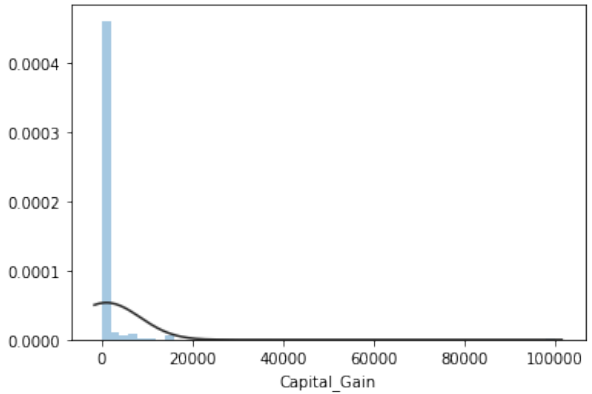


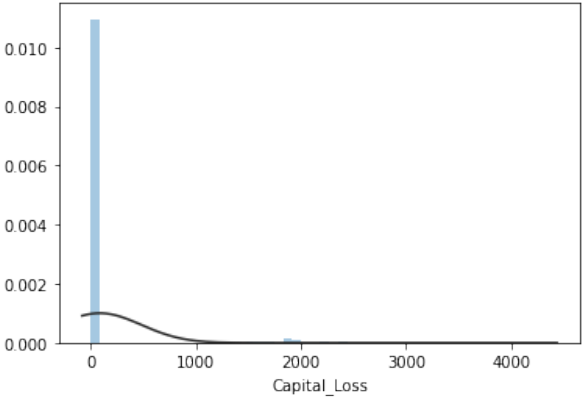
Occupation v.s. Income

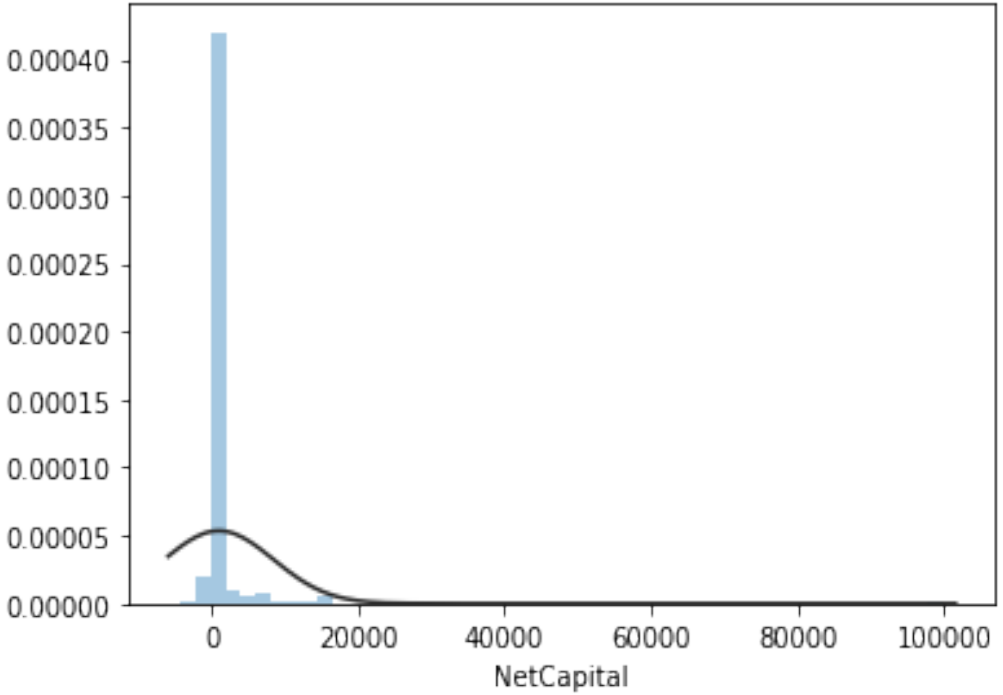


Data Distribution of Independent Variables

Net_Capital = Capital_Gain - Capital_Loss

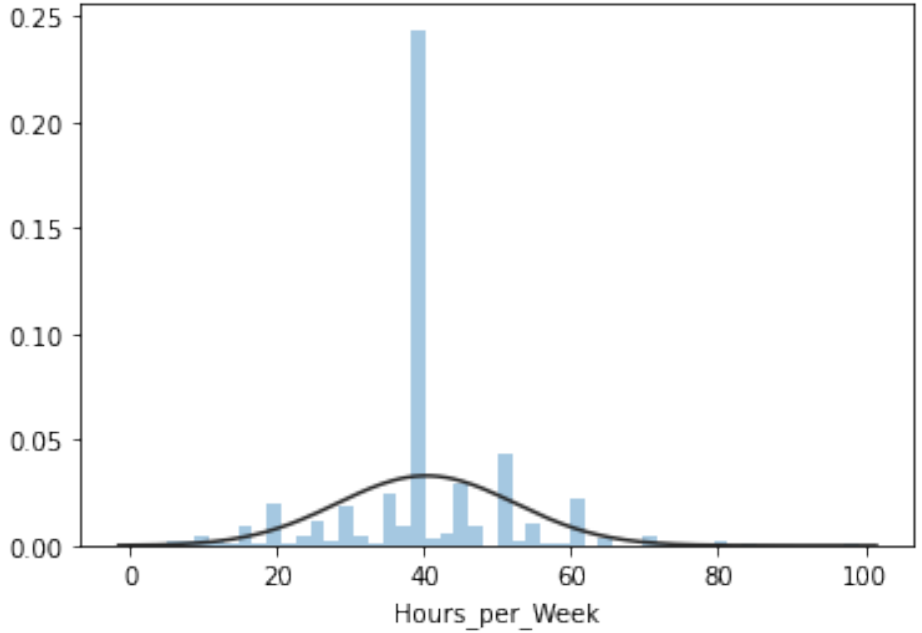






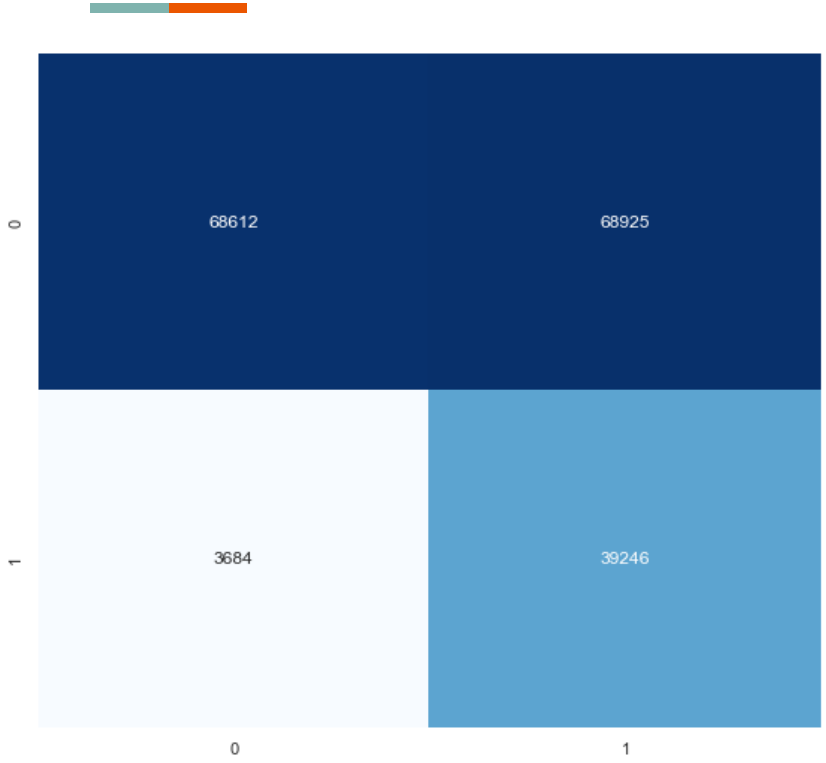
Data Distribution of Independent Variables

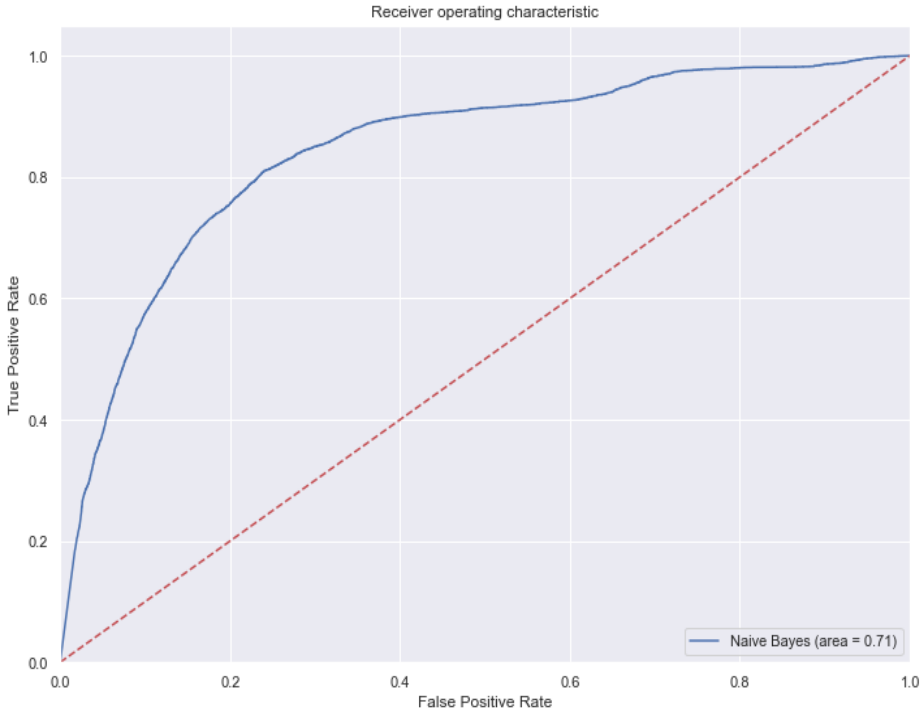
Hours Per Week



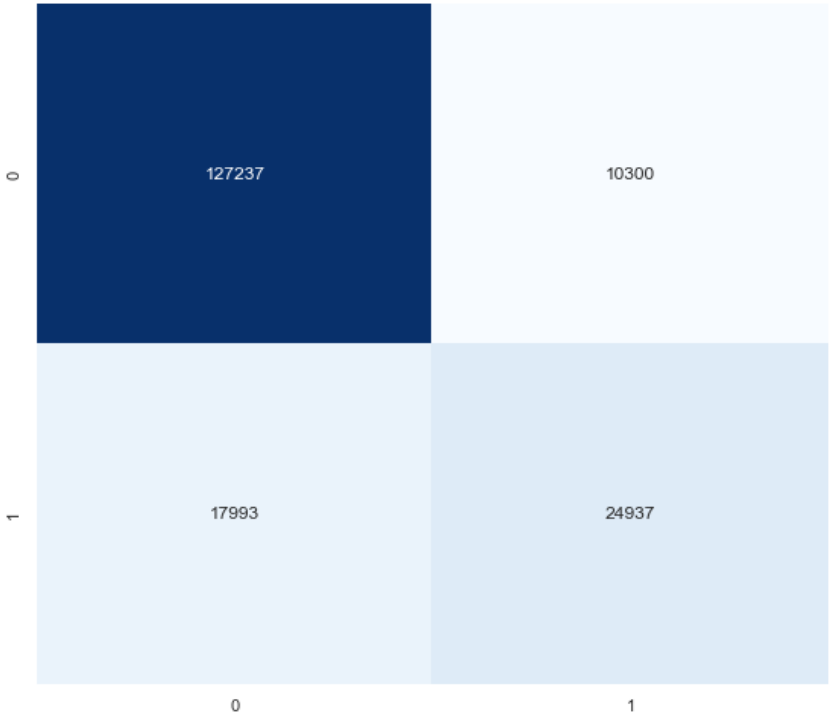
Outliers found in Hours Per Week. We imputed outliers with interpolation nearest method.

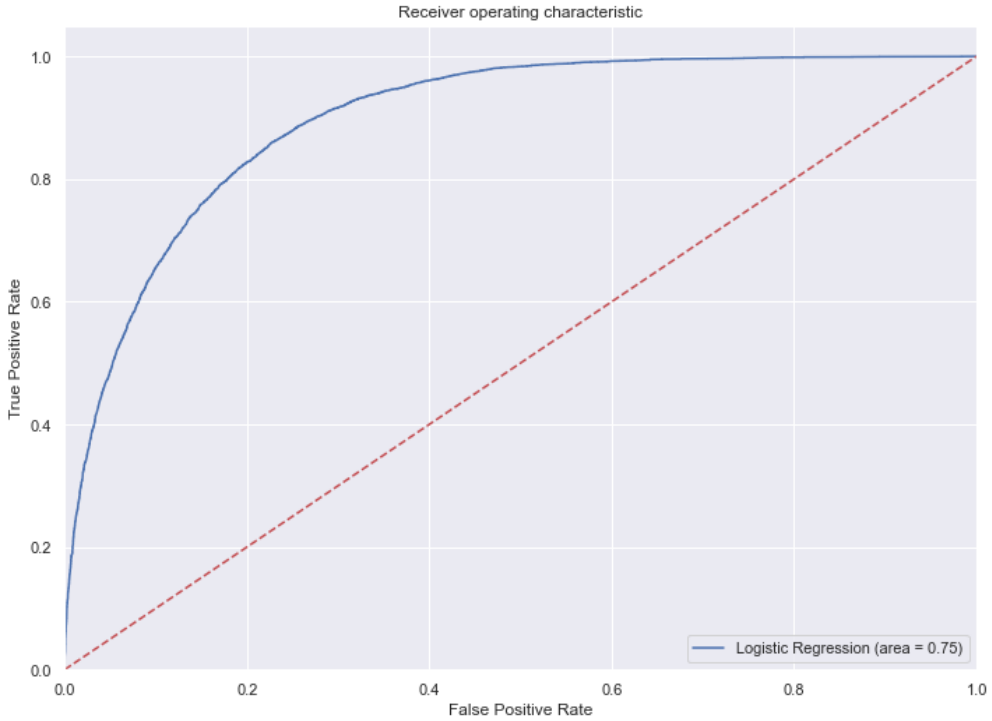
Naive Bayes Confusion Matrix & ROC Curve



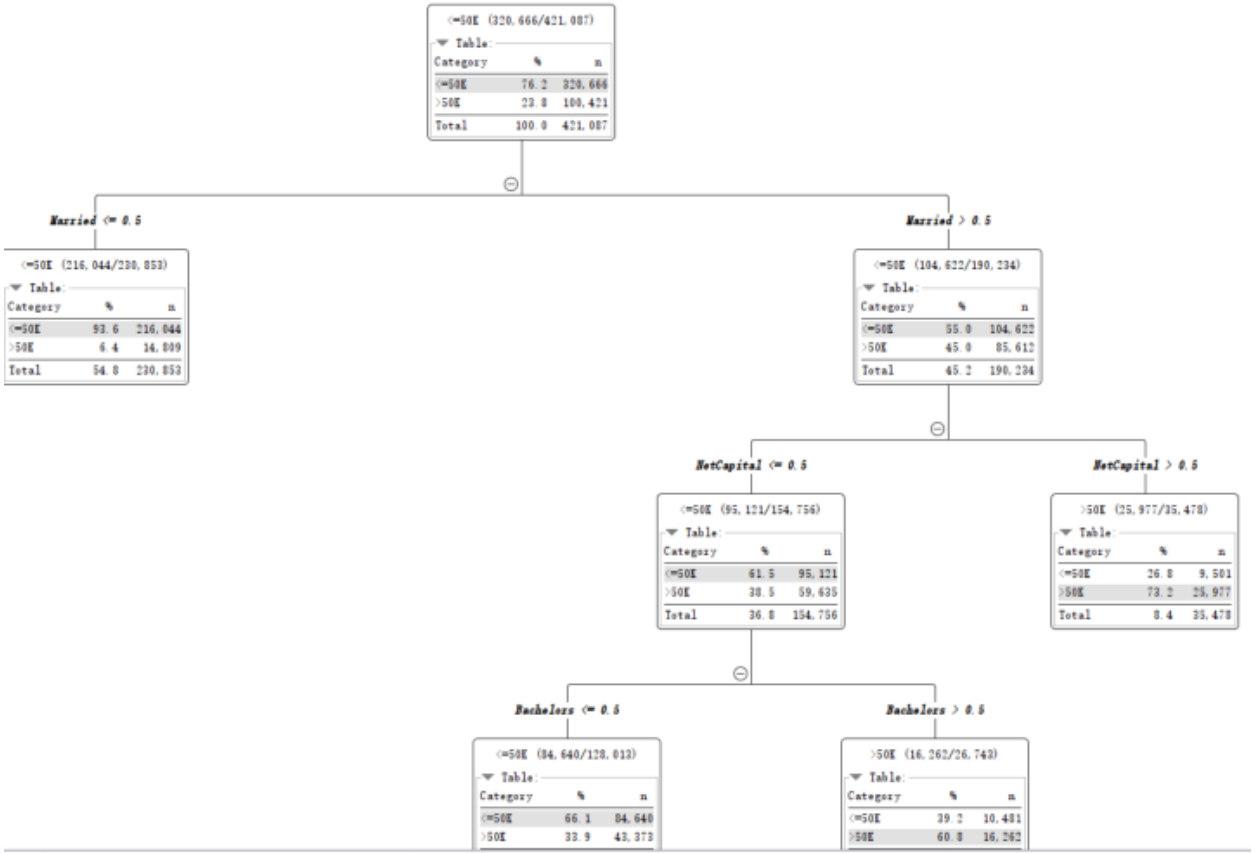


Logistic Regression Confusion Matrix & ROC Curve

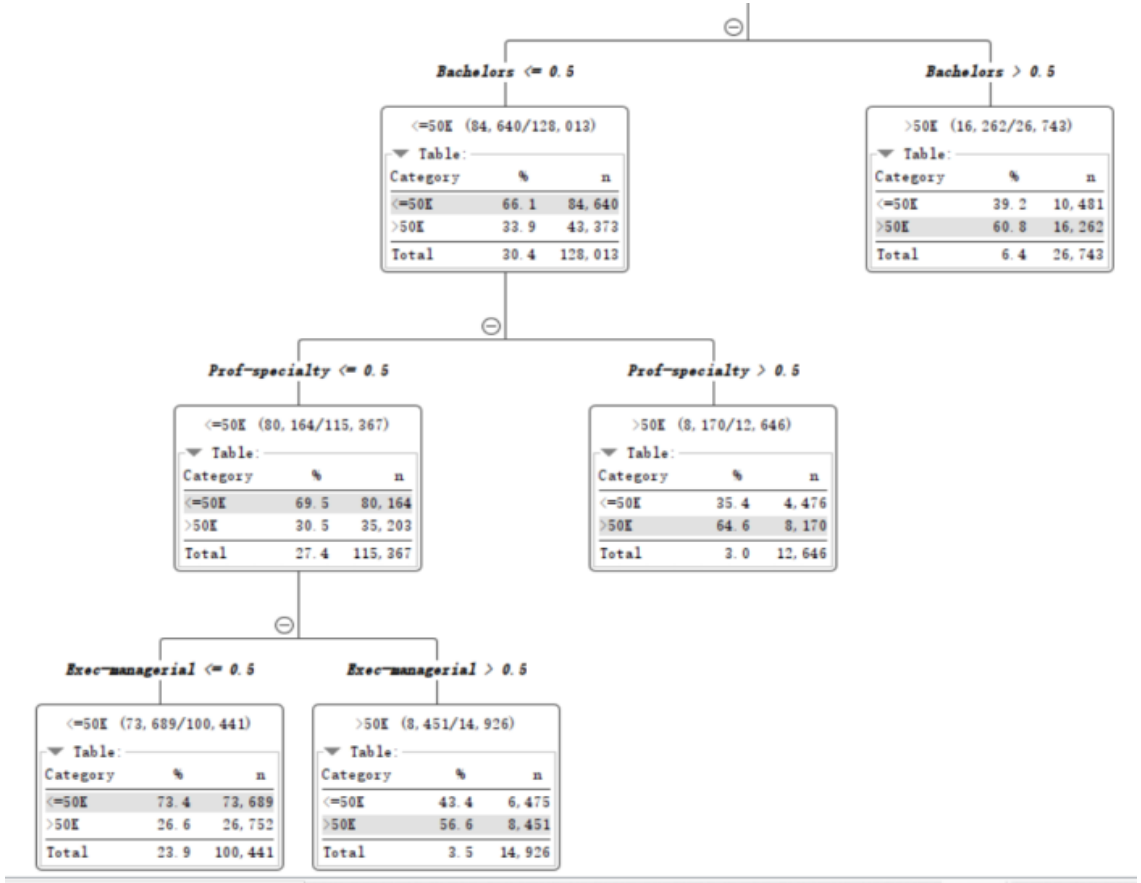




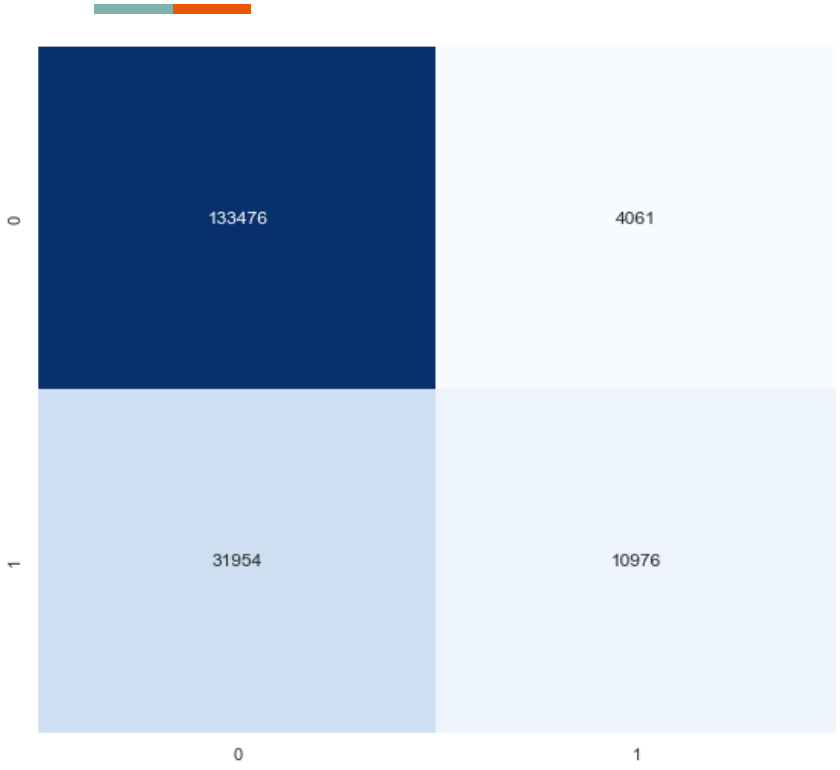
Decision Tree

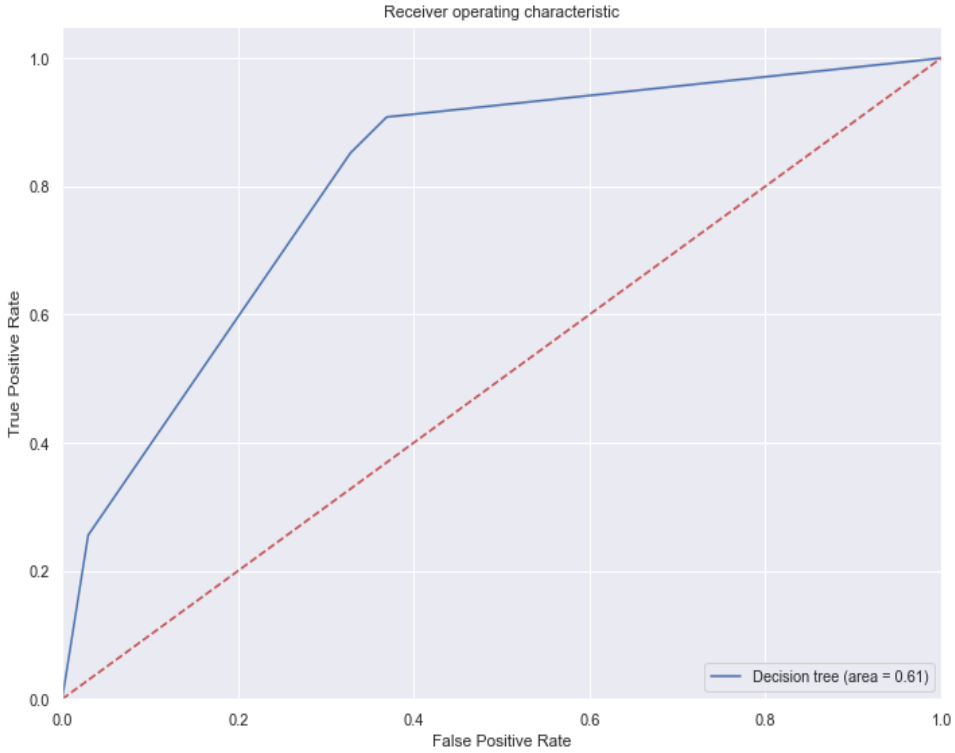


Decision Tree (cont.)

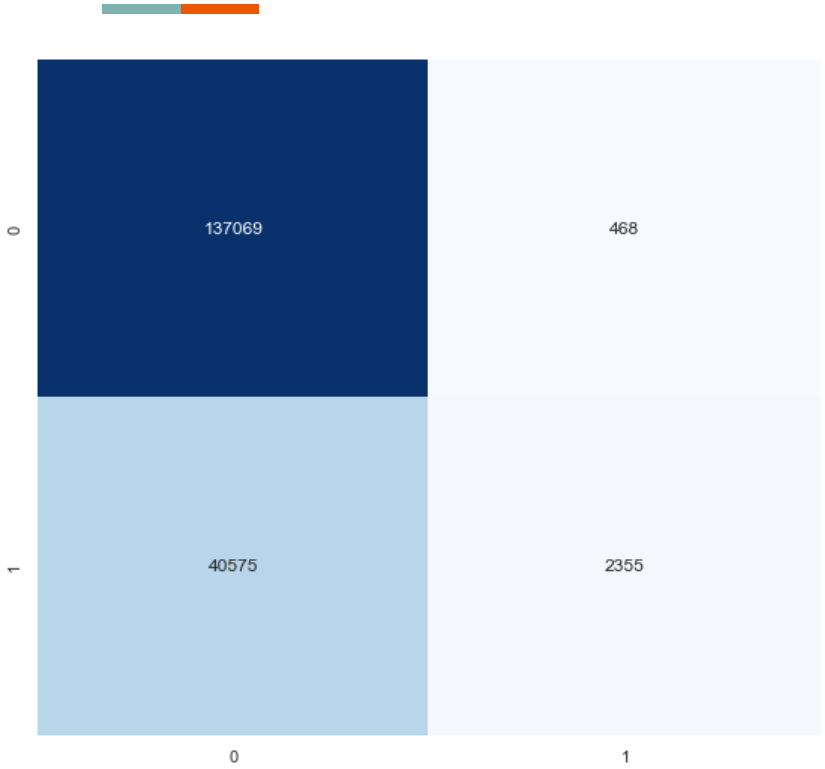


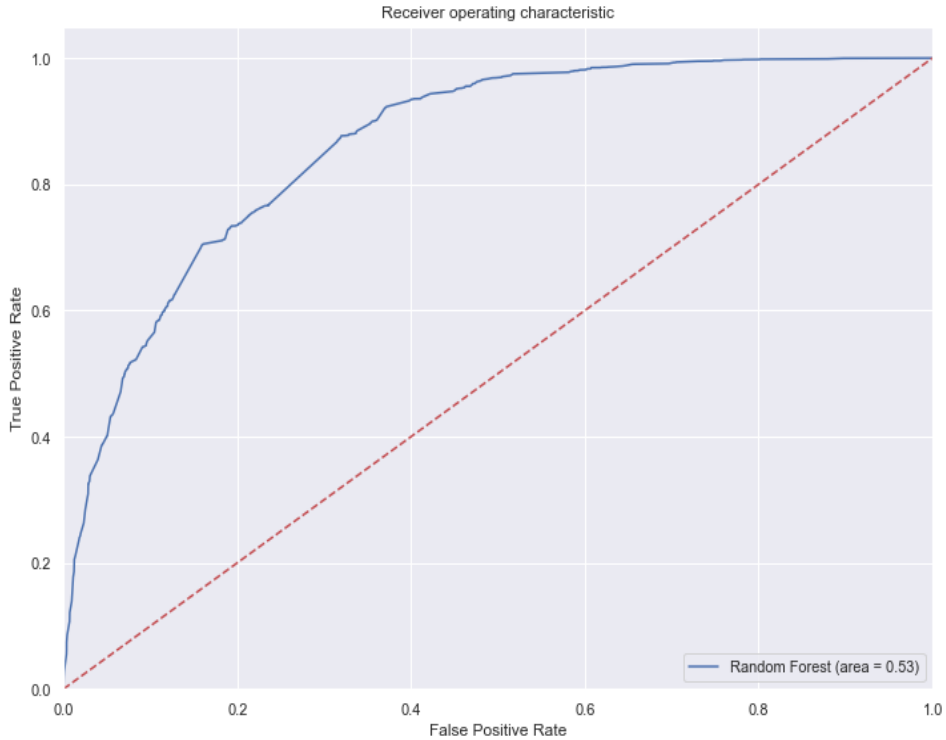
Decision Tree Confusion Matrix & ROC Curve





Random Forest Confusion Matrix & ROC Curve





Statistical Summary of Top Numeric Predictors

Results: Logit

Model:	Logit	Pseudo R-squared:	0.390
Dependent Variable:	Income	AIC:	282298.1276
Date:	2021-11-10 19:25	BIC:	282758.0526
No. Observations:	421087	Log-Likelihood:	-1.4111e+05
Df Model:	41	LL-Null:	-2.3145e+05
Df Residuals:	421045	LLR p-value:	0.0000
Converged:	0.0000	Scale:	1.0000
No. Iterations:	35.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Age	0.0265	0.0004	60.8013	0.0000	0.0256	0.0273
Hours_per_Week	0.0389	0.0005	76.0321	0.0000	0.0379	0.0399
NetCapital	1.4974	0.0125	119.5144	0.0000	1.4729	1.5220