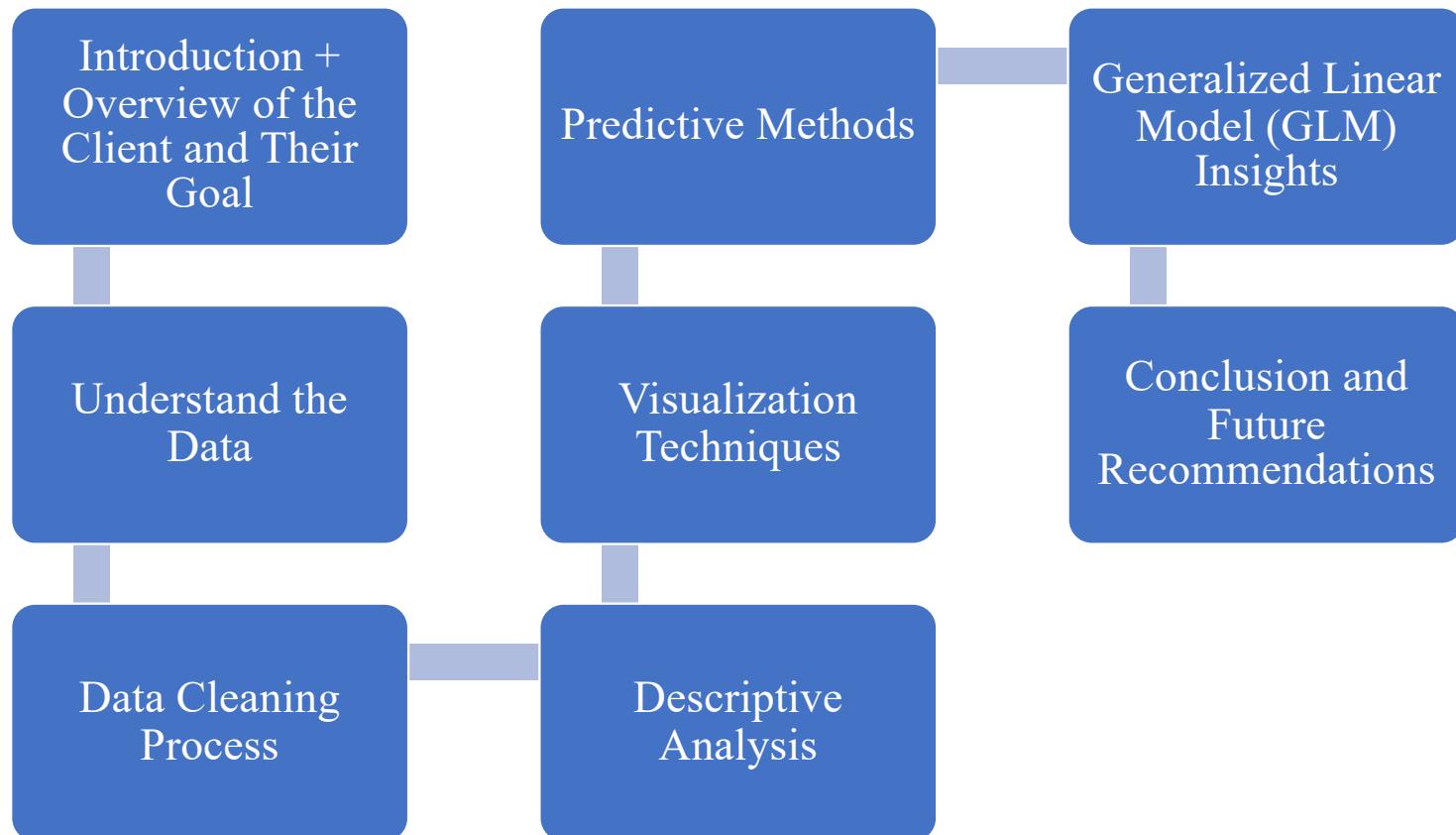




Enhancing Skincare Recommendations for The Glow Company

By Chi Dang

Agenda:





Data Introduction + Client's Overview & Goals

- Client: The Glow Company
- Question: What demographic and employment factors have the strongest correlation with high income status (over \$50K/year)? Identifying key predictors.
- Focus/ Goals: What influencing factors that The Glow Company's consumers have income above \$50K
- Collaboration Objective: Refining skincare recommendations for a more effective and personalized routine

Understand the Data

- Dataset: Adult Income Census from Kaggle
- Origin: Extracted from 1994 Census Bureau by Ronny Kohavi and Barry Becker.
- Number of Rows: 32,561
- Number of Columns: 15
- Y variable: Binary outcome indicating income >\$50K
- X variables: Age, Marital_Status, Race, Sex, Hrs_per_Week, Income, Net_Capital, EducationLevel_clean, Occupation_clean, & WorkingClass_clean.
- Tools using in this project: R Studio & SAS



Data Cleaning Process

- Handling Special Characters, like '?' in the dataset
- Rename columns for a purpose of using it easily in both R & SAS
- Adding a new column Net_Capital by subtracting Net_Gain to Net_Loss
- Transform the irrelevant variables into one group
- Using R Studio to finish the cleaning process and extracting file to computer to upload on SAS
- Transform Y variables into binary variable, 0 as <=50K & 1 as >50K

Before Cleaning:

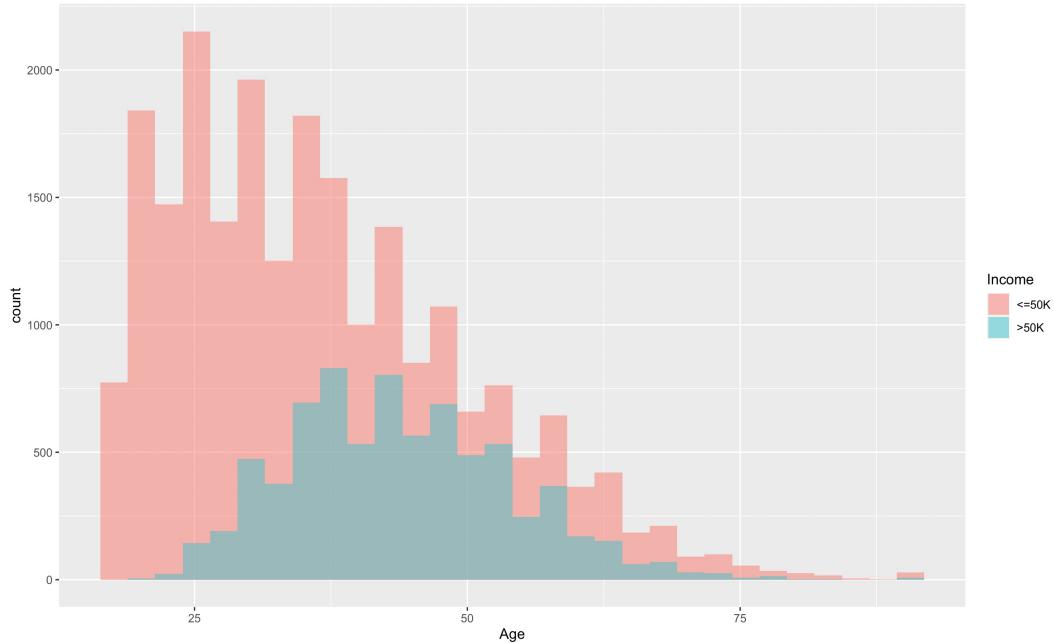
```
> str(adult)    # look at the structure of the data
'data.frame': 32561 obs. of 15 variables:
 $ age          : int 90 82 66 54 41 34 38 74 68 41 ...
 $ workclass    : Factor w/ 9 levels "?","Federal-gov",...: 1 5 1 5 5 5 5 8 2 5 ...
 $ fnlwgt       : int 77053 132870 186061 140359 264663 216864 150601 88638 422013 70037 ...
 $ education    : Factor w/ 16 levels "10th","11th",...: 12 12 16 6 16 12 1 11 12 16 ...
 $ education.num: int 9 9 10 4 10 9 6 16 9 10 ...
 $ marital.status: Factor w/ 7 levels "Divorced","Married-AF-spouse",...: 7 7 7 1 6 1 6 5 1 5 ...
 $ occupation   : Factor w/ 15 levels "?","Adm-clerical",...: 1 5 1 8 11 9 2 11 11 4 ...
 $ relationship : Factor w/ 6 levels "Husband","Not-in-family",...: 2 2 5 5 4 5 5 3 2 5 ...
 $ race         : Factor w/ 5 levels "Amer-Indian-Eskimo",...: 5 5 3 5 5 5 5 5 5 ...
 $ sex          : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 2 1 1 2 ...
 $ capital.gain : int 0 0 0 0 0 0 0 0 0 ...
 $ capital.loss : int 4356 4356 4356 3900 3900 3770 3770 3683 3683 3004 ...
 $ hours.per.week: int 40 18 40 40 45 40 40 40 60 ...
 $ native.country: Factor w/ 42 levels "?","Cambodia",...: 40 40 40 40 40 40 40 40 40 1 ...
 $ income        : Factor w/ 2 levels "<=50K",>50K": 1 1 1 1 1 1 2 1 2 ...
```

After Cleaning:

```
> str(adult2)
'data.frame': 30162 obs. of 10 variables:
 $ Age          : int 82 54 41 34 38 74 68 45 38 52 ...
 $ Marital_Status: Factor w/ 7 levels "Divorced","Married-AF-spouse",...: 7 1 6 1 6 5 1 1 5 7 ...
 $ Race         : Factor w/ 5 levels "Amer-Indian-Eskimo",...: 5 5 5 5 5 5 3 5 5 ...
 $ Sex          : Factor w/ 2 levels "Female","Male": 1 1 1 2 1 1 1 2 1 ...
 $ Hrs_per_Week : int 18 40 40 45 40 20 40 35 45 20 ...
 $ Income        : Factor w/ 2 levels "<=50K",>50K": 1 1 1 1 1 2 1 2 2 2 ...
 $ Net_Capital  : int -4356 -3900 -3900 -3770 -3770 -3683 -3683 -3004 -2824 -2824 ...
 $ EducationLevel_clean: chr "HighSchool" "MiddleSchool" "Post-High_School" "HighSchool" ...
 $ Occupation_clean : chr "Professional" "SkilledLabor" "Professional" "Services" ...
 $ WorkingClass_clean : chr "Private" "Private" "Private" "Private" ...
```

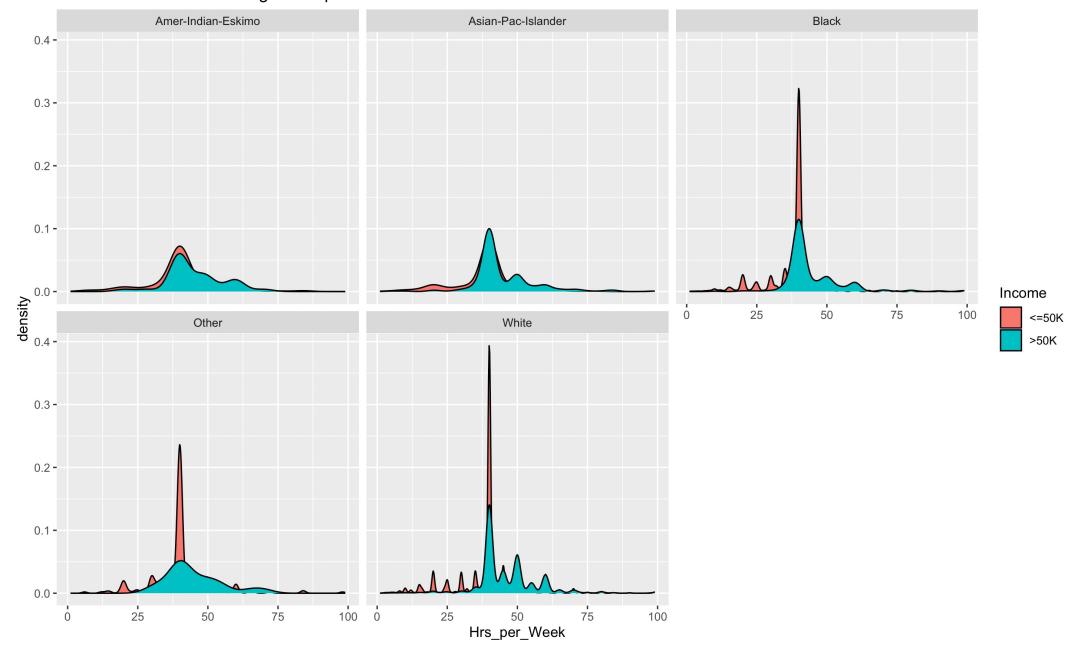
Visualization

Age Distribution by Income Level



Income
≤50K
>50K

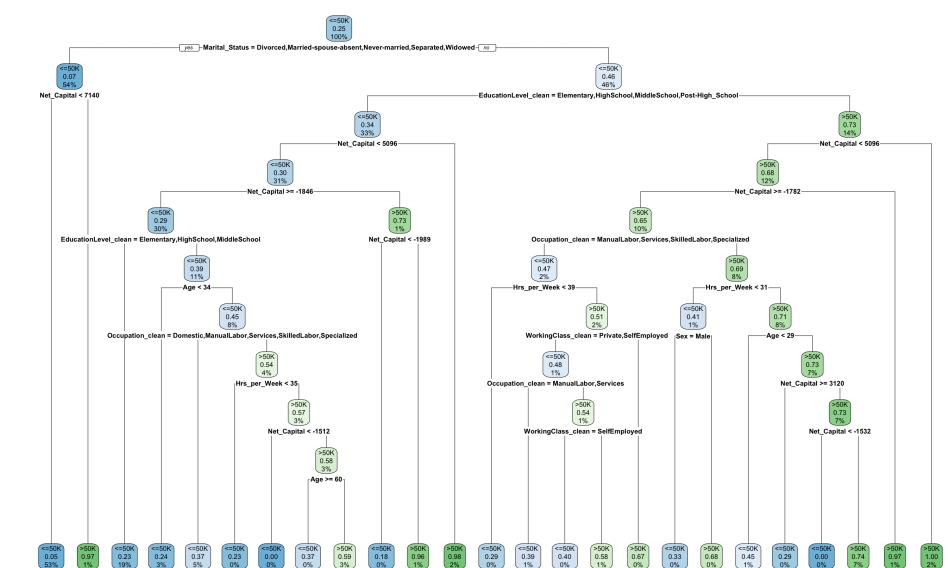
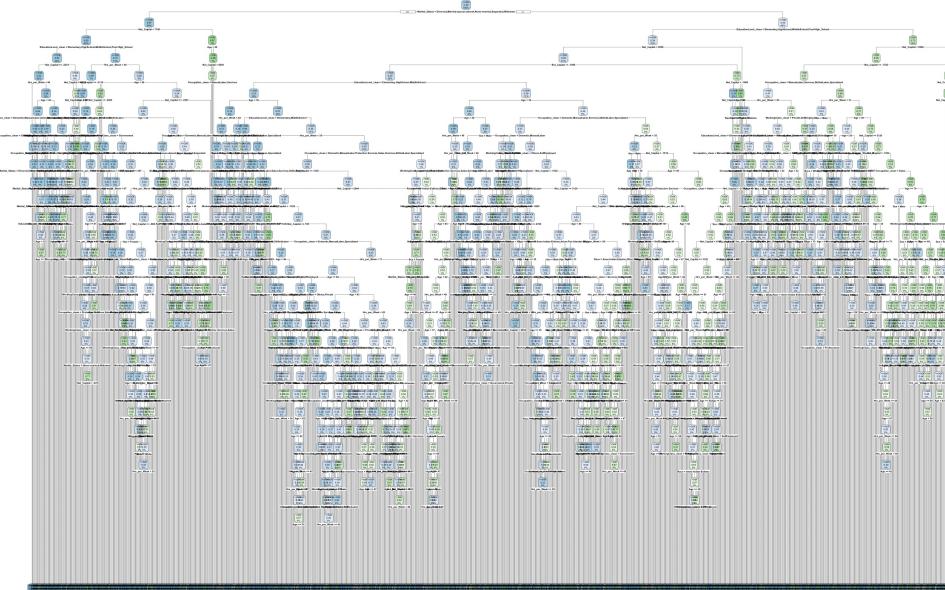
Income of different Working Hours per Week from different race

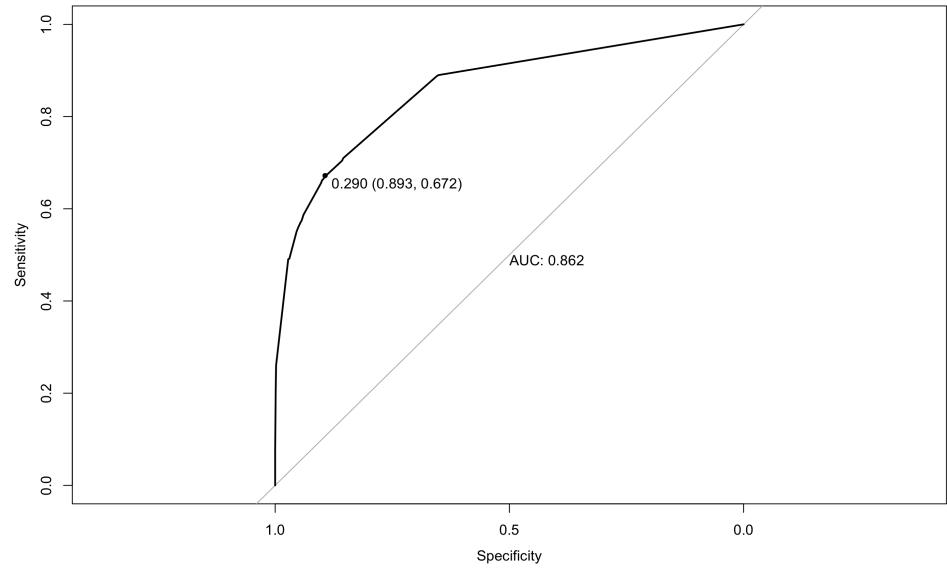
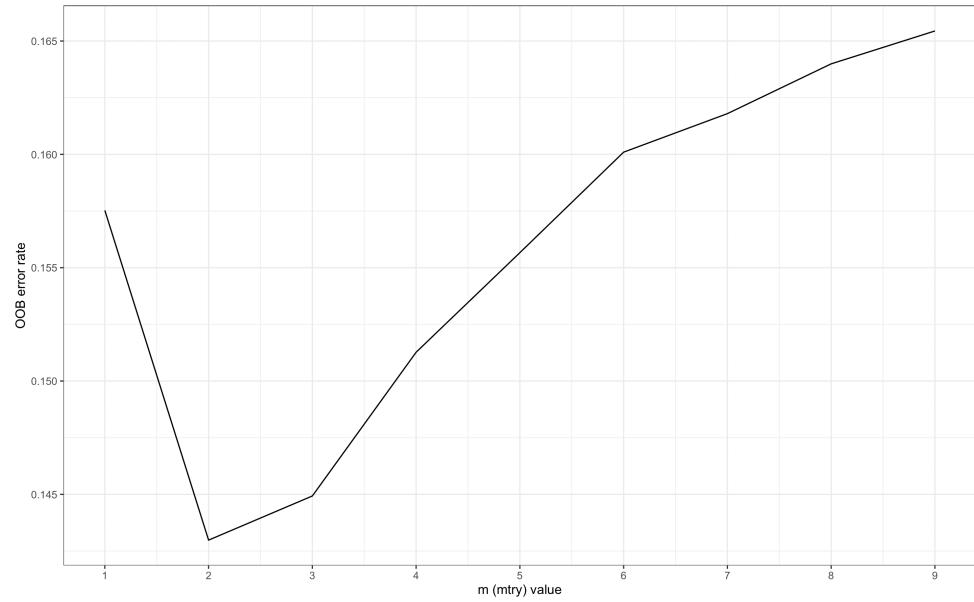


Income
≤50K
>50K

Predictive Methods Pt.1 (Decision Tree)

Before and After Tuning the tree

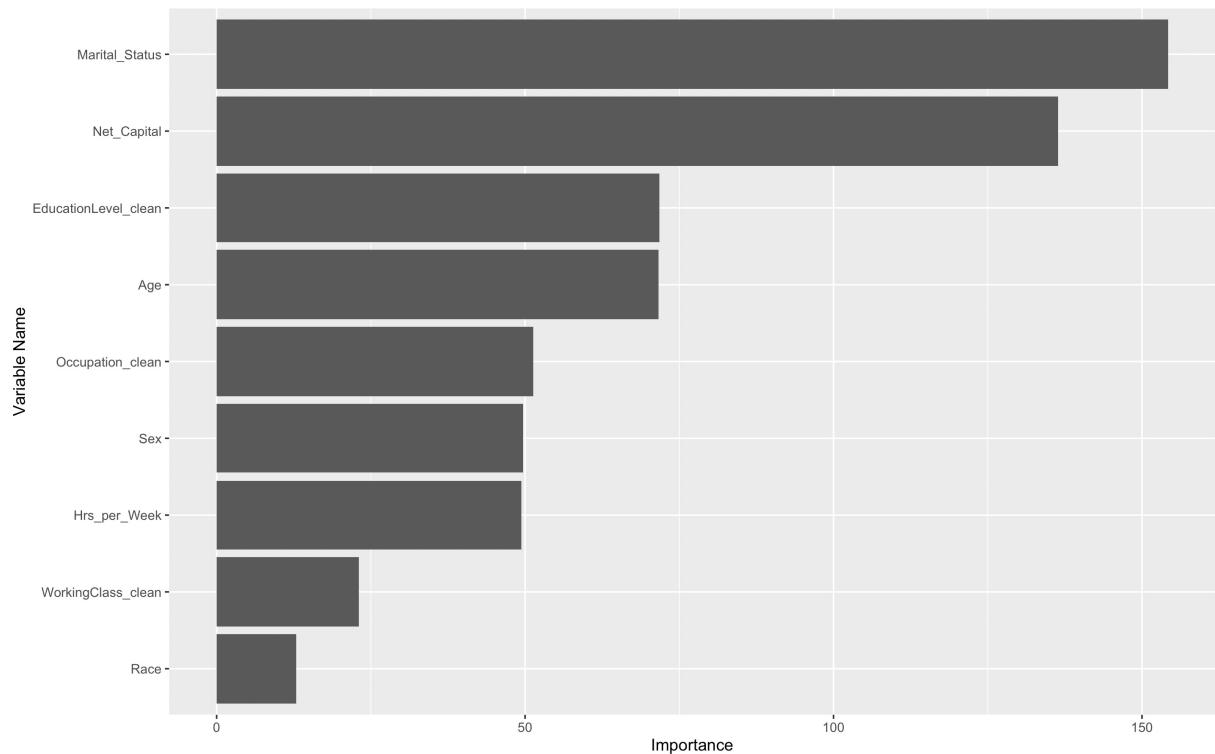




Predictive Method Pt.2 (Random Forest)

Useful plots for creating GLM

- The graph shows us that Martial_Status is the most important variable that affects an individual's income to be more than >\$50K.
- However, I took a further step to understand the relationship between income and other variables can affect a person's earning.
- And, all of variables affects a person's incoming by fitting the model and comparing their AIC/BIC, which the m8 includes all of variables in has the smallest AIC.



GLM Descriptive Methods (Bernoulli)

- $Y \sim \text{Bernoulli}(\pi_i)$
 - $g(\pi_i) = \eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$
 - $\eta_i = -5.4880 + 0.0279 * (\text{Age}) + 0.3179 * (\text{WorkingClass_Clean_Government}) - 1.9377 * (\text{WorkingClass_clean_Others}) + 0.3281 * (\text{WorkingClass_clean_Private}) + 0.000245 * (\text{Net_Capital}) - 0.679 * (\text{Race_Amer-In}) - 0.1146 * (\text{Race_Amer-P}) - 0.154 * (\text{Race_Black}) - 0.7889 * (\text{Race_Other}) + 2.1163 * (\text{EducationLevel_clean_Elementary}) - 0.5298 * (\text{EducationLevel_clean_HighSchool}) + 0.8685 * (\text{EducationLevel_clean_HigherEducation}) - 1.8588 * (\text{EducationLevel_clean_MiddleSchool}) - 0.1691 * (\text{Sex_Female}) + 0.0298 * (\text{Hrs_per_Week}) + 0.00982 * (\text{Marital_Status_Divorced}) + 2.9706 * (\text{Marital_Status_Married-AF-spouse}) + 2.1275 * (\text{Marital_Status_Married-civ-spouse}) - 0.0792 * (\text{Marital_Status_Married-spouse-absen}) - 0.5231 * (\text{Marital_Status_Never-married}) - 0.0887 * (\text{Marital_Status_Separated}) - 3.4447 * (\text{Occupation_clean_Domestic}) - 0.1809 * (\text{Occupation_clean_ManualLabor}) + 0.8786 * (\text{Occupation_clean_Professional}) + 0.5492 * (\text{Occupation_clean_Protective Ser}) + 0.4277 * (\text{Occupation_clean_Sales}) - 0.1114 * (\text{Occupation_clean_Services}) + 0.0739 * (\text{Occupation_clean_SkilledLabor})$

Analysis of Penalized Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-5.4880	0.2084	693.6576	<.0001
Age		1	0.0279	0.00160	304.3733	<.0001
WorkingClass_clean	Governemnt	1	0.3179	0.0661	23.1022	<.0001
WorkingClass_clean	Others	1	-1.9377	1.5799	1.5042	0.2200
WorkingClass_clean	Private	1	0.3281	0.0515	40.6720	<.0001
Net_Capital		1	0.000245	8.638E-6	803.1012	<.0001
Race	Amer-In	1	-0.6790	0.2193	9.5863	0.0020
Race	Asian-P	1	-0.1146	0.1006	1.2961	0.2549
Race	Black	1	-0.1540	0.0729	4.4594	0.0347
Race	Other	1	-0.7899	0.2809	7.9105	0.0049
EducationLevel_clean	Elementary	1	-2.1163	0.2564	68.1085	<.0001
EducationLevel_clean	HighSchool	1	-0.5298	0.0442	143.6988	<.0001
EducationLevel_clean	HigherEducation	1	0.8685	0.0471	339.5380	<.0001
EducationLevel_clean	MiddleSchool	1	-1.8588	0.1462	161.7114	<.0001
Sex	Female	1	-0.1691	0.0517	10.7148	0.0011
Hrs_per_Week		1	0.0298	0.00158	352.6052	<.0001
Marital_Status	Divorced	1	0.00982	0.1485	0.0044	0.9473
Marital_Status	Married-AF-spouse	1	2.9743	0.5130	33.6142	<.0001
Marital_Status	Married-civ-spouse	1	2.1275	0.1428	221.8993	<.0001
Marital_Status	Married-spouse-absen	1	-0.0792	0.2580	0.0943	0.7588
Marital_Status	Never-married	1	-0.5231	0.1515	11.9142	0.0006
Marital_Status	Separated	1	-0.0887	0.2004	0.1961	0.6579
Occupation_clean	Domestic	1	-3.4447	1.2394	7.7244	0.0054
Occupation_clean	ManualLabor	1	-0.1809	0.0932	3.7640	0.0524
Occupation_clean	Professional	1	0.8786	0.0762	133.0843	<.0001
Occupation_clean	Protective Ser	1	0.5492	0.1260	18.9848	<.0001
Occupation_clean	Sales	1	0.4277	0.0837	26.1002	<.0001
Occupation_clean	Services	1	-0.1114	0.0861	1.6754	0.1955
Occupation_clean	SkilledLabor	1	0.0739	0.0790	0.8769	0.3490

Coefficient Interpretations



95% confident interval for the odds ratio corresponding to a female is $(e^{[-0.2705]}, e^{[-0.0679]}) = (0.7629, 0.9343)$. It means that we are 95% confident that the odds of a female's income is more than \$50k is increasing from 6.6% to 23.7%.



95% confident interval for the odds ratio corresponding to a 20-year-old is $(e^{[3*(0.0247)}], e^{[3*(0.0310)]}) = (1.0769, 1.097)$. It means that we are 95% confident that the odds of a 20-year-old's income is more than \$50k is increasing from 7.7% to 9.7%.



For a year increase in age, an individual's odds of having income more than \$50K change by a factor of $e^{(0.279)} = 1.0289$, hold other factors constant.

Conclusion:



The project's consideration of a wide range of demographic variables as potential predictors, from marital status to race demonstrates a comprehensive approach to understanding income determinants.



Strengthened Predictive Models by using both decision tree and random forest, so that I can use



Aiming to refine analysis and align with The Glow Company's mission for a more effective and personalized skincare routine.



If you have any questions, please don't hesitate to reach out to me via chi.dang@drake.edu