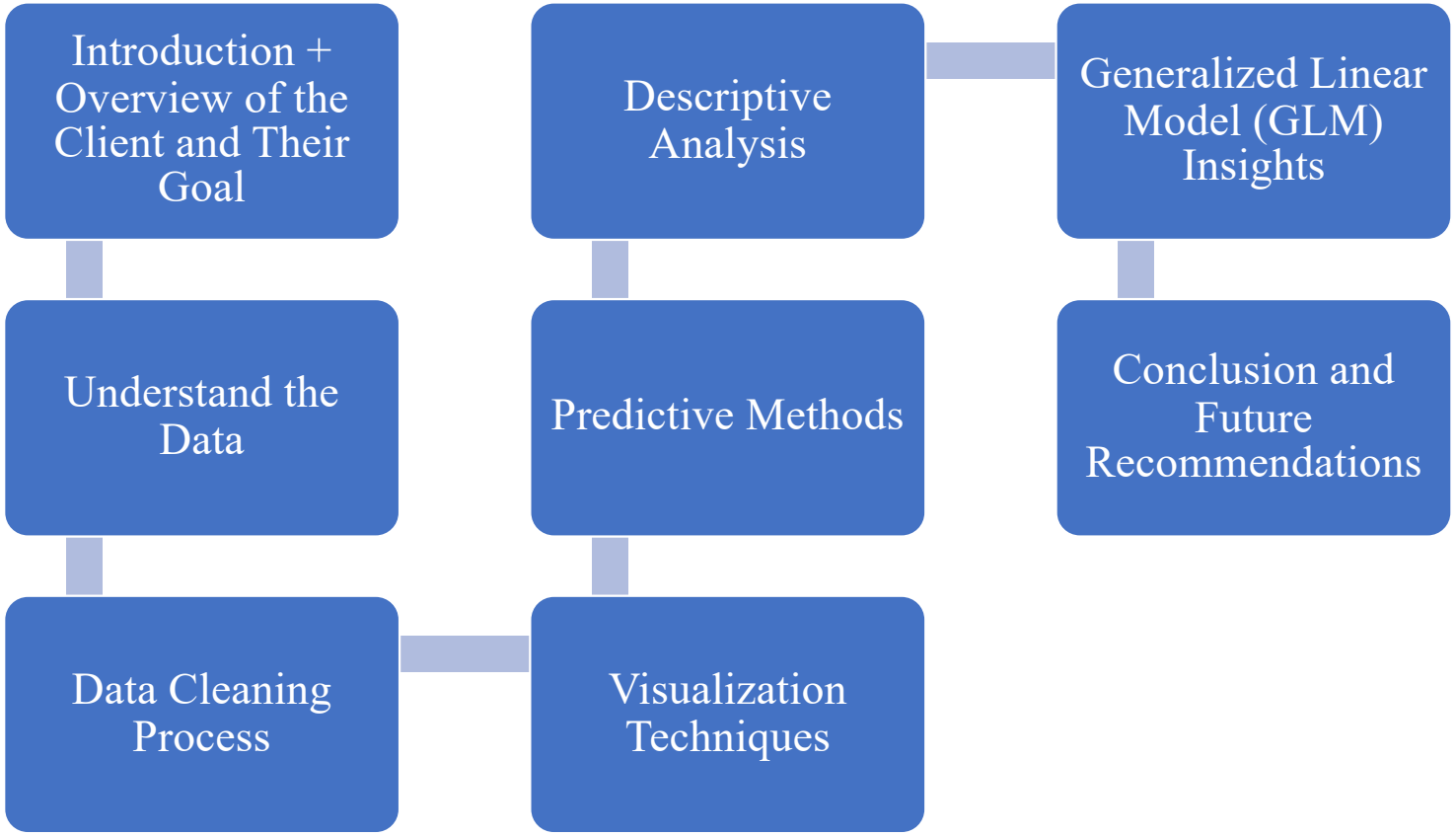

Enhancing Skincare Recommendations for The Glow Company
By Chi Dang

Agenda:


Introduction + Overview of the Client and Their Goal
Understand the Data
Data Cleaning Process
Visualization Techniques
Predictive Methods
Descriptive Analysis
Generalized Linear Model (GLM) Insights
Conclusion and Future Recommendations

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

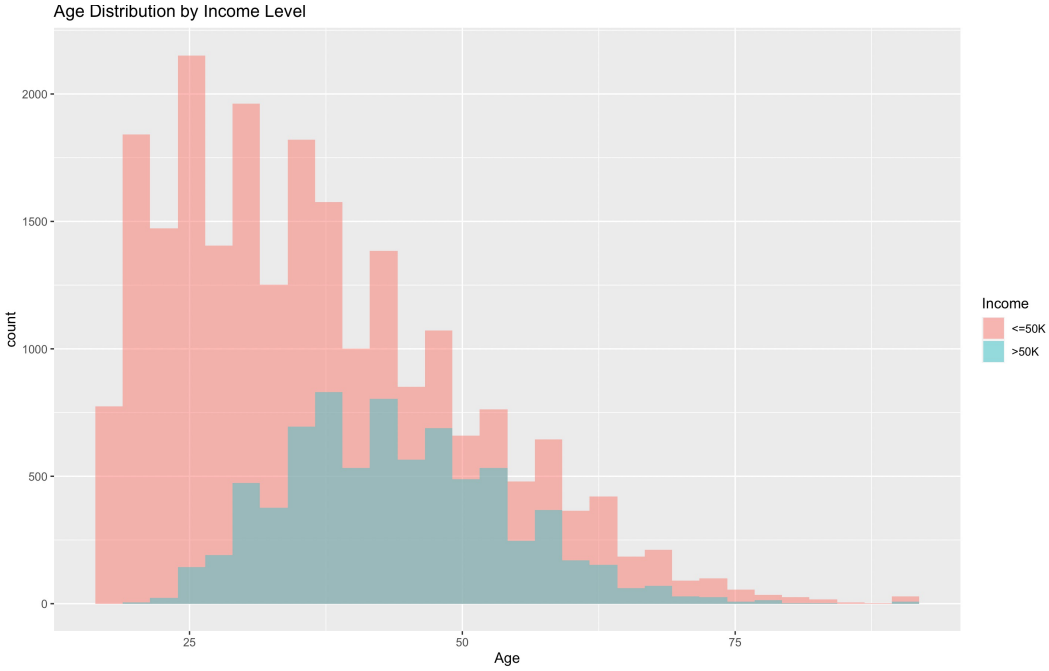
Before Cleaning:

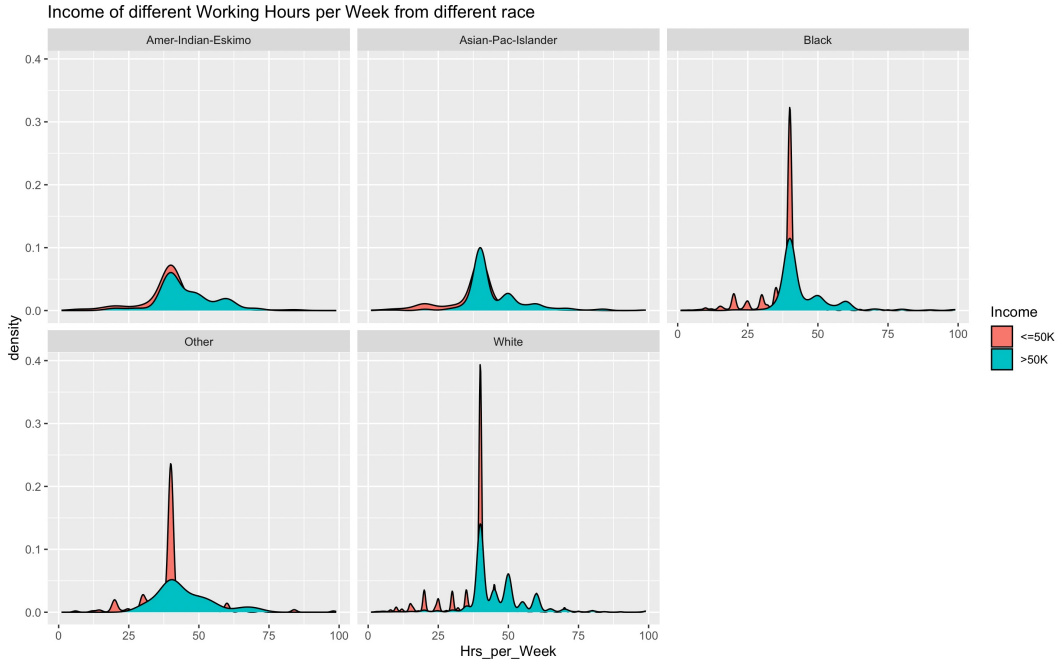
```
> str(adult)        # look at the structure of the data
'data.frame':   32561 obs. of  15 variables:
 $ age           : int  90 82 66 54 41 34 38 74 68 41 ...
 $ workclass     : Factor w/ 9 levels "?","Federal-gov",..: 1 5 1 5 5 5 5 8 2 5 ...
 $ fnlwgt        : int  77053 132870 186061 140359 264663 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 5 ...
 $ sex           : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 2 1 1 2 ...
 $ capital.gain  : int  0 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 40 45 40 20 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 1 2 1 2 ...
```

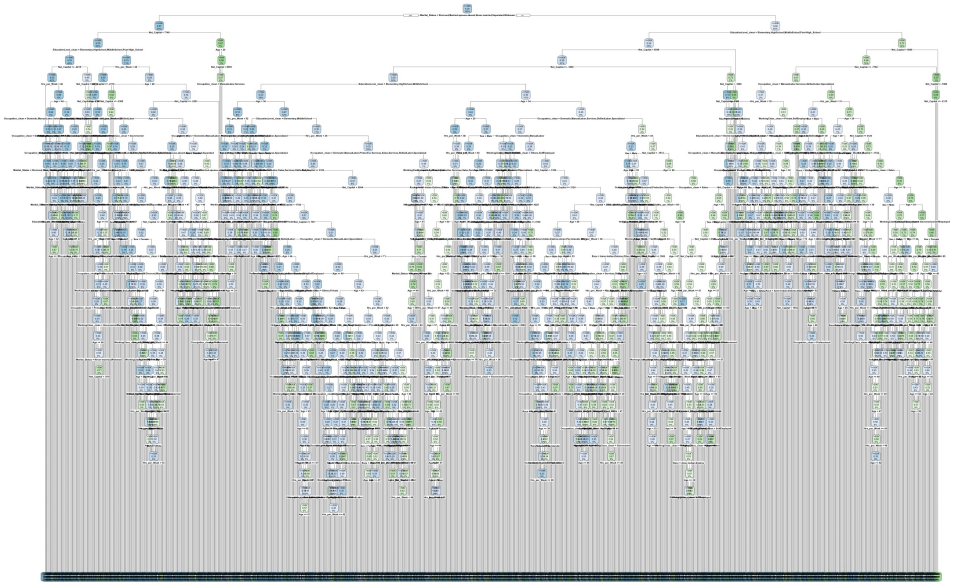
After Cleaning:

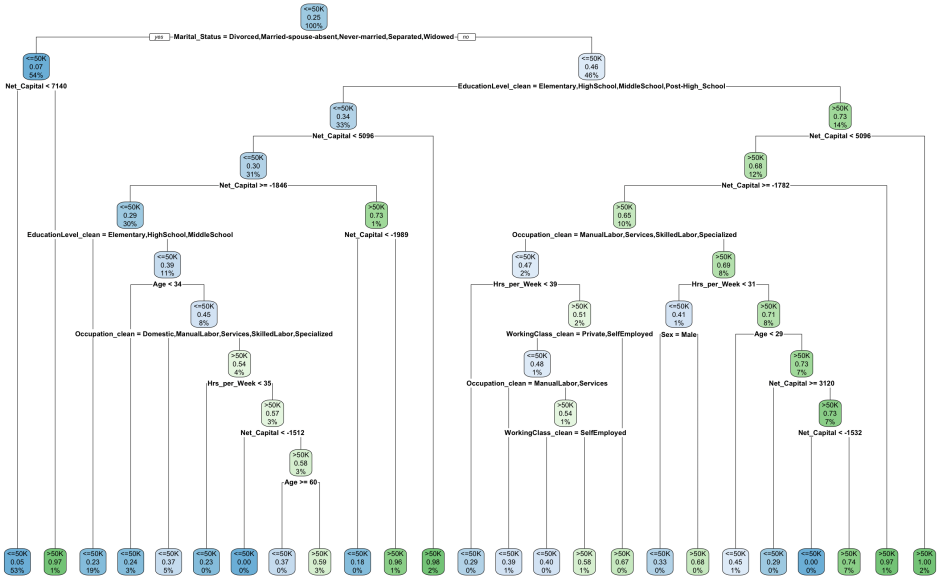
```
> str(adult2)
'data.frame':   30162 obs. of  10 variables:
 $ Age                 : int  82 54 41 34 38 74 68 45 38 52 ...
 $ Marital_Status      : Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 7 1 6 1 6 5 1 1 5 7 ...
 $ Race                : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 5 5 5 5 5 3 5 5 ...
 $ Sex                 : Factor w/ 2 levels "Female","Male": 1 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







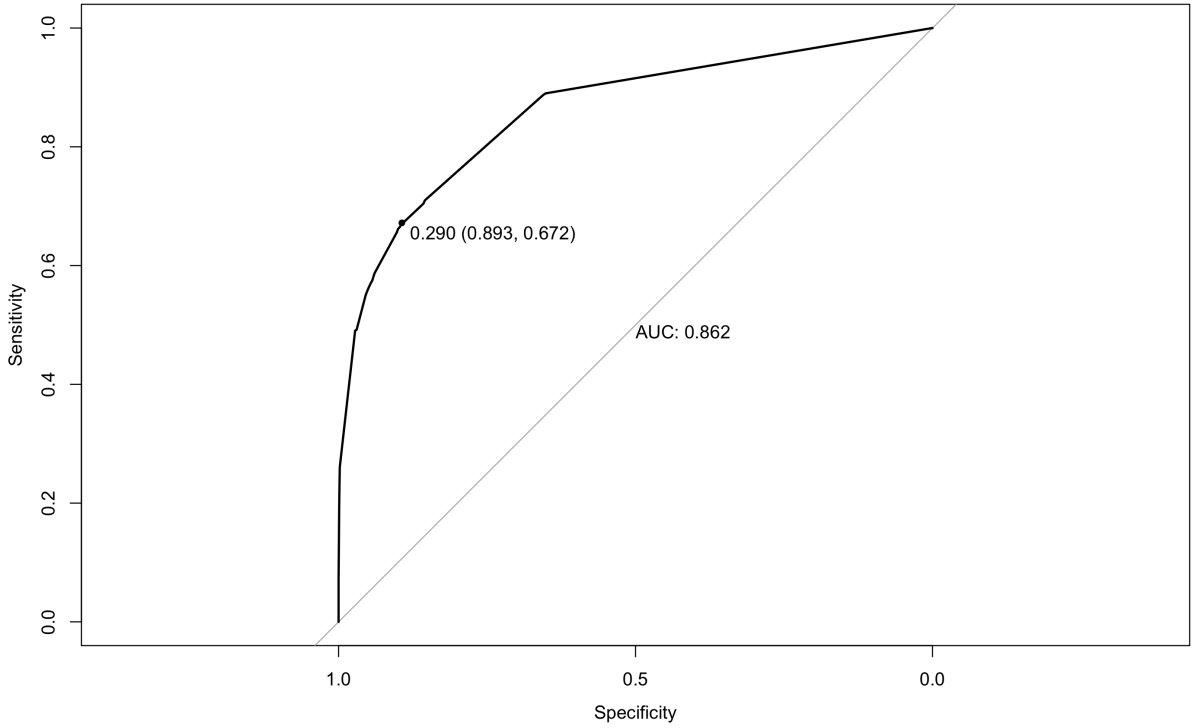


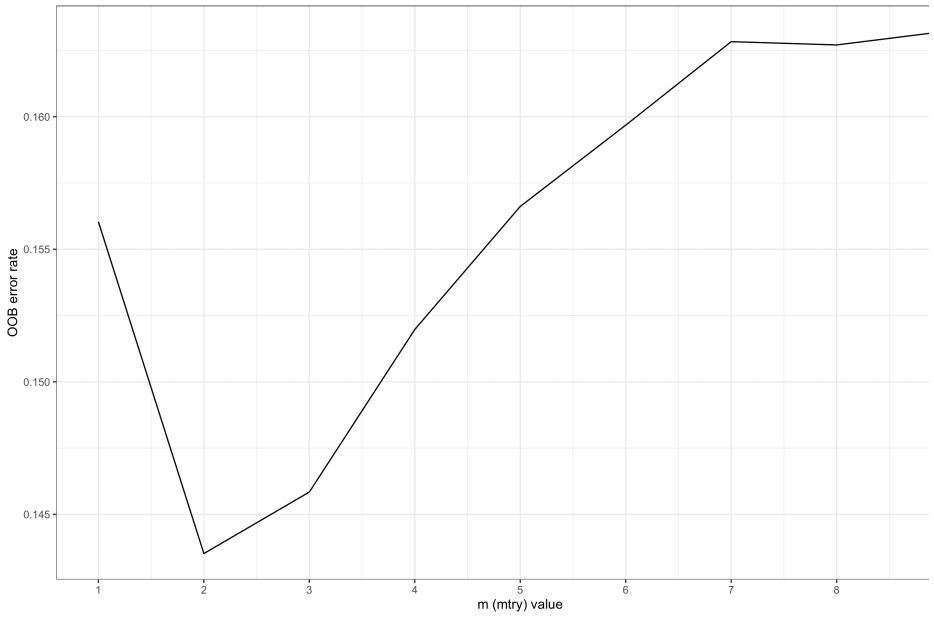
Predictive Methods Pt.1
(Decision Tree Pt.1)

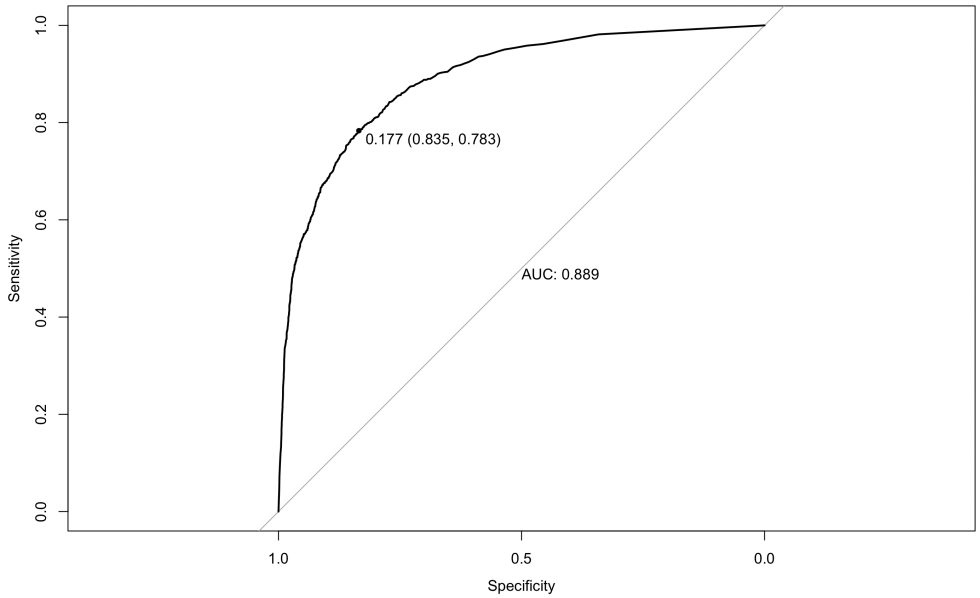
Before and After Tunning the tree

Predictive Methods Pt.2 (Decision Tree Pt.2)

- We will predict 89.3% of the time when an individual's earning is less than $50K and 67.7% of the time when an individual's earning is more than $50K.
- Area Under the Curve is 0.862, which is decent but we will see if the random forest AUC is higher than this one.





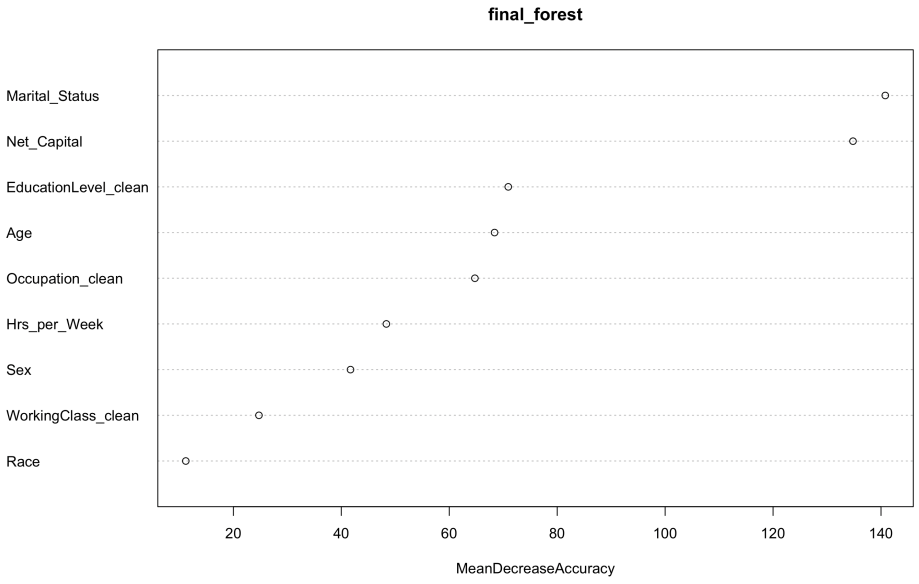


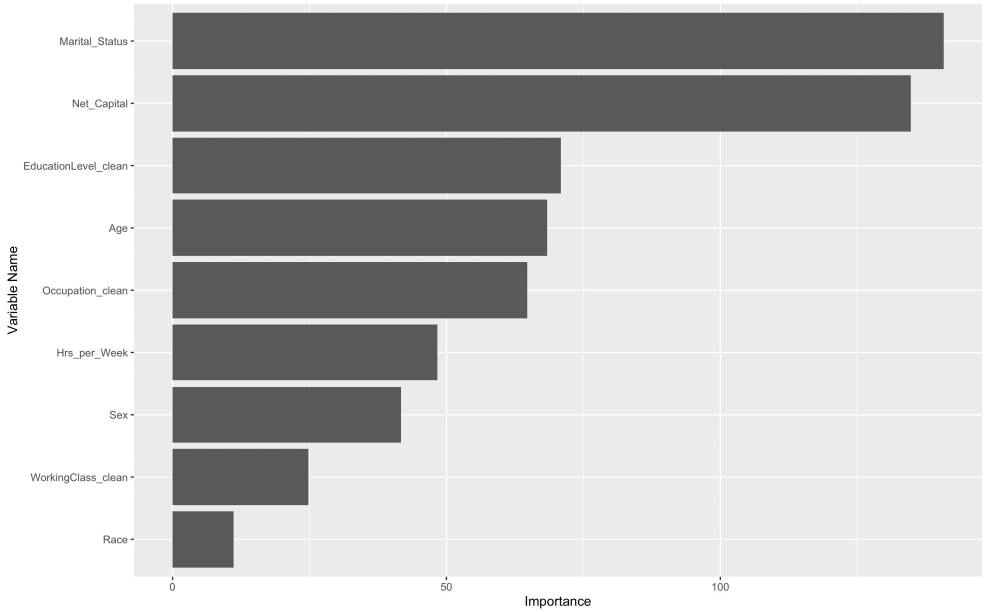
Predictive Method Pt.3 (Random Forest)

- We will predict 83.5% of the time when an individual's earning is less than $50K and 78.3% of the time when an individual's earning is more than $50K.
- Area Under the Curve is 0.889

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 m9 includes all of variables in has the smallest AIC, 21084.48.





GLM Descriptive Methods (Bernoulli)

- Y ~ Bernoulli (π_i)
  - g (π_i) = $\eta_i = \log(\frac{\pi_i}{1-\pi_i})$

η_i = -5.4985 + 0.0124*(Marital_Status_Divorced) + 2.9706*(Marital_Status_Married-AF-spouse) + 2.1332*(Marital_Status_Married-civ-spouse) - 0.0912*(Marital_Status_Married-spouse-absen) - 0.5209*(Marital_Status_Never-married) - 0.0926*(Marital_Status_Separated) + 0.000246*(Net_Capital) - 2.1450*(EducationLevel_clean_Elementary) - 0.5305*(EducationLevel_clean_HighSchool) + 0.8698*(EducationLevel_clean_HigherEducation) - 1.8678*(EducationLevel_clean_MiddleSchool) + 0.0279*(Age) - 3.7302*(Occupation_clean_Domestic) - 0.1808*(Occupation_clean_Manual) + 0.8802*(Occupation_clean_Profess) + 0.5498*(Occupation_clean_Protective Ser) + 0.4286*(Occupation_clean_Sales) - 0.1113*(Occupation_clean_Services) + 0.0748*(Occupation_clean_SkilledLabor) + 0.0298*(Hrs_per_Week) - 0.1695*(Sex_Female) + 0.3182*(WorkingClass_Clean_Government) - 10.7283*(WorkingClass_clean_Others) + 0.3286*(WorkingClass_clean_Private) - 0.6914*(Race_Amer-In) - 0.1158*(Race_ Amer-P) - 0.1551*(Race_Black) - 0.8115*(Race_other)

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-5.4985	0.2087	694.0229	<.0001
Marital_Status	Divorced	1	0.0124	0.1488	0.0070	0.9334
Marital_Status	Married-AF-spouse	1	2.9706	0.5135	33.4699	<.0001
Marital_Status	Married-civ-spouse	1	2.1332	0.1431	222.0979	<.0001
Marital_Status	Married-spouse-absen	1	-0.0912	0.2594	0.1236	0.7252
Marital_Status	Never-married	1	-0.5209	0.1519	11.7629	0.0006
Marital_Status	Separated	1	-0.0926	0.2010	0.2121	0.6452
Net_Capital		1	0.000246	8.654E-6	805.1891	<.0001
EducationLevel_clean	Elementary	1	-2.1450	0.2596	68.2631	<.0001
EducationLevel_clean	HighSchool	1	-0.5305	0.0442	143.8690	<.0001
EducationLevel_clean	HigherEducation	1	0.8698	0.0472	340.1415	<.0001
EducationLevel_clean	MiddleSchool	1	-1.8678	0.1467	162.1131	<.0001
Age		1	0.0279	0.00160	304.4554	<.0001
Occupation_clean	Domestic	1	-3.7302	1.3053	8.1664	0.0043
Occupation_clean	ManualLabor	1	-0.1808	0.0933	3.7549	0.0527
Occupation_clean	Professional	1	0.8802	0.0762	133.2963	<.0001
Occupation_clean	Protective Ser	1	0.5498	0.1261	19.0006	<.0001
Occupation_clean	Sales	1	0.4286	0.0838	26.1539	<.0001
Occupation_clean	Services	1	-0.1113	0.0862	1.6675	0.1966
Occupation_clean	SkilledLabor	1	0.0748	0.0790	0.8947	0.3442
Hrs_per_Week		1	0.0298	0.00159	352.9813	<.0001
Sex	Female	1	-0.1695	0.0517	10.7372	0.0011
WorkingClass_clean	Governemnt	1	0.3182	0.0662	23.1172	<.0001
WorkingClass_clean	Others	1	-10.7283	119.0	0.0081	0.9281
WorkingClass_clean	Private	1	0.3286	0.0515	40.7362	<.0001
Race	Amer-In	1	-0.6914	0.2201	9.8654	0.0017
Race	Asian-P	1	-0.1158	0.1007	1.3217	0.2503
Race	Black	1	-0.1551	0.0730	4.5140	0.0336
Race	Other	1	-0.8115	0.2828	8.2320	0.0041

Coefficient Interpretations

95% confident interval for the odds ratio corresponding to a female is (e^[(- 0.2709)], e^[(-0.0682)]) = (0.7629, 0.934). It means that we are 95% confident that the odds of a female's income is more than $50k is decreasing from 6.6% to 23.7%.

95% confident interval for the odds ratio corresponding to a 20-year-old is (e^[3*(0.0248)], 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 3-year increase in age, an individual's odds of having income more than $50K change by a factor of e^(3*0.279) = 1.08, 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.

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

Strengthened Predictive Models by using both decision tree and random forest, to see which model predicts the data better and decision tree has a larger AUC which I will choose it over random forest

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