STAT 448 - Final Project Report

Team 2 - Kashyap Ava, Sunny Chen

Introduction

Our project aims to predict whether an individual's annual income exceeds $50,000 based on 11 socio-economic and salary-related features. The original dataset from the UCI Machine Learning Repository (Barry Becker did an extraction from the 1994 Census database) has almost 50,000 observations and 14 features. Still, after sampling a portion of this dataset and preprocessing the data, we are left with 917 observations and 11 features along with our response variable, salary. Throughout this report, we will address the following questions: First, we will present a general descriptive summary of the socio-economic and salary-related features considered in this study, including interesting descriptive analyses, tables, and plots for both continuous and categorical variables to provide a clear overview of the data, while considering the removal of records associated with categorical variables that have few observations for certain levels. Second, we will conduct an appropriate statistical analysis to test the potential association between salary categories and each predictor, whether categorical (nominal/ordinal) or continuous, and draw main conclusions from these preliminary analyses.

Descriptive Analysis & Potential Associations:

The dataset we are working with has been preprocessed to have 11 socio-economic and salary-related features, and our response is salary. We will dive into an interesting descriptive analysis of each feature and create tables and plots for each feature. There are three continuous variables: Age, Ednum and Hpw. The remaining predictors are categorical variables out of which only Education is the only ordinal variable, and the rest are nominal.

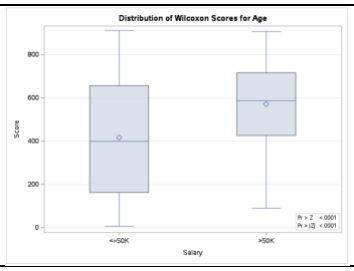
We identified an observation with WC as “Without-pay” and removed it as it is obvious that the salary is less than 50k. Then, we regrouped the Education variable with groups “Pre-school”, “1st-4th”, “5th-6th”, “7th-8th”, “9th”, “10th”, “11th” and “12th” into a new group called “School” and then the groups of “Assoc-acdm” and “Assoc-voc” into a new group and named it “Associate”. This was done to ensure they affect both the categories of the target variable salary so that any model can capture the effect appropriately. Similarly, we grouped the MS variable with groups “Married-AF-spouse”, “Married-civ-spouse” and “Married-spouse-absent” into one group called “Married”. Additionally, the Race variable with groups “Amer-Indian-Eskimo” and “Asian-Pac-Islander” into a new group called “Eskimo-Islander”. These regrouping were done to ensure that the groups do not belong to a single pure class but have values in both the classes to potentially extract the effect of the categorical variable. We observed that there were three observations with NCountry value being “South”. These were removed as the native country information is ambiguous from the label “South”. The NCountry was then regrouped into “US” and “Non-US” categories as most of the observations are from the US, we thought it would be a better inference to look at it from that perspective. The following discussion entails the descriptive analysis of the variables and their association with the target variable salary. We used the Likelihood ratio chi-squared tests for association between the nominal categorical variables and salary, Mantel-Haenszel chi-squared test for association between Education (ordinal categorical variable) and Wilcoxon rank sum tests

for the association between the continuous variables and salary. Th chi-squared tests were used instead of the Fischer's exact test because the cell counts are large, and the Wilcoxon rank sum test was used instead of the t-test as the normality assumptions were not met by the continuous predictors.

Age

Basic Statistical Measures			
Location		Variability	
Mean	39.10624	Std Deviation	13.44819
Median	37.00000	Variance	180.85383
Mode	36.00000	Range	73.00000
		Interquartile Range	20.00000

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.07598405	Pr > D	<0.010
Cramer-von Mises	W-Sq	1.01097578	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	6.53842003	Pr > A-Sq	<0.005



Descriptive Statistics	Tests for Normality	Association with Salary

Inference: The age variable is not normally distributed and hence Wilcoxon Rank sum test was used. The test concludes that there is a clear difference between the levels of salary based on age and hence can be interpreted as a significant predictor for the target salary.

Ednum

Basic Statistical Measures			
Location		Variability	
Mean	10.09748	Std Deviation	2.56416
Median	10.00000	Variance	6.57492
Mode	9.00000	Range	15.00000
		Interquartile Range	4.00000

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.2149360	Pr > D	<0.010
Cramer-von Mises	W-Sq	6.9340600	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	35.7017836	Pr > A-Sq	<0.005

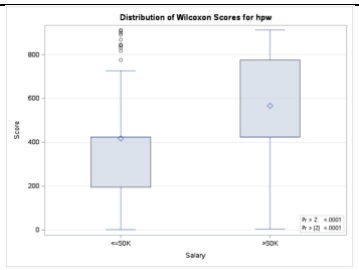
Descriptive Statistics	Tests for Normality	Association with Salary

Inference: The Ednum variable is not normally distributed and hence Wilcoxon Rank sum test was used. The test concludes that there is a clear difference between the levels of salary based on Ednum and hence can be interpreted as a significant predictor for the target salary.

Hpw

Basic Statistical Measures			
Location		Variability	
Mean	40.56627	Std Deviation	11.64752
Median	40.00000	Variance	135.66474
Mode	40.00000	Range	95.00000
		Interquartile Range	5.00000

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.2549826	Pr > D	<0.010
Cramer-von Mises	W-Sq	11.7072628	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	53.0295554	Pr > A-Sq	<0.005



Descriptive Statistics	Tests for Normality	Association with Salary

Inference: The Hpw variable is not normally distributed and hence Wilcoxon Rank sum test was used. The test concludes that there is a clear difference between the levels of salary based on Hpw and hence can be interpreted as a significant predictor (may not be strongly) for the target salary.

Education_new

Education_new	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Associate	57	6.24	57	6.24
Bachelors	160	17.52	217	23.77
Doctorate	7	0.77	224	24.53
HS-grad	313	34.28	537	58.82
Masters	50	5.48	587	64.29
Prof-schoo	18	1.97	605	66.27
School	109	11.94	714	78.20
Some-colle	199	21.80	913	100.00

Descriptive Statistics

Table of Education_new by Salary

Education_new	Salary <=50K	Salary >50K	Total
Associate	39 42.079	18 14.921	57
Bachelors	97 118.12	63 41.884	160
Doctorate	1 5.1676	6 1.8324	7
HS-grad	257 231.06	56 81.935	313
Masters	20 36.911	30 13.089	50
Prof-schoo	6 13.288	12 4.7119	18
School	105 80.467	4 28.533	109
Some-colle	149 146.91	50 52.093	199
Total	674	239	913

Table of Education_new by Salary

Statistic	DF	Value	Prob
Chi-Square	7	112.7978	<.0001
Likelihood Ratio Chi-Square	7	115.4877	<.0001
Mantel-Haenszel Chi-Square	1	11.8963	0.0006
Phi Coefficient		0.3515	
Contingency Coefficient		0.3316	
Cramer's V		0.3515	

Tests for Association

Inference: The Education_new is somewhat evenly distributed amongst the levels and has 8 levels. The Mantel-Haenszel chi-squared test (preferred as high cell counts and ordinal) concludes that there is a significant association between the target salary and Education_new variable and hence can be interpreted as a significant predictor for the target salary.

WC

WC	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Federal-gov	34	3.72	34	3.72
Local-gov	68	7.45	102	11.17
Private	659	72.18	761	83.35
Self-emp-inc	27	2.96	788	86.31
Self-emp-not-inc	82	8.98	870	95.29
State-gov	43	4.71	913	100.00

Descriptive Statistics

Table of WC by Salary

WC	Salary <=50K	Salary >50K	Total
Federal-gov	20 25.1	14 8.9003	34
Local-gov	47 50.199	21 17.801	68
Private	500 486.49	159 172.51	659
Self-emp-inc	12 19.932	15 7.0679	27
Self-emp-not-inc	63 60.535	19 21.465	82
State-gov	32 31.744	11 11.256	43
Total	674	239	913

Table of WC by Salary

Statistic	DF	Value	Prob
Chi-Square	5	18.6202	0.0023
Likelihood Ratio Chi-Square	5	16.6065	0.0053
Mantel-Haenszel Chi-Square	1	0.7091	0.3997
Phi Coefficient		0.1428	
Contingency Coefficient		0.1414	
Cramer's V		0.1428	

Tests for Association

Inference: The WC is dominated by “Private” class and has 6 levels. The likelihood ratio chi-squared test (preferred as high cell counts) concludes that there is a significant association between the target salary and WC variable and hence can be interpreted as a significant predictor for the target salary.

MS_new

Inference: The MS_new is somewhat evenly distributed amongst the levels and has 5 levels. The likelihood ratio chi-squared test (preferred as high cell counts) concludes that there is a significant association between the target salary and MS_new variable and hence can be interpreted as a significant predictor for the target salary.

MS_new	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Divorced	122	13.36	122	13.36
Married	453	49.62	575	62.98
Never-marr	292	31.98	867	94.96
Separated	21	2.30	888	97.26
Widowed	25	2.74	913	100.00

Descriptive Statistics

Table of MS_new by Salary

MS_new	Salary		
	<=50K	>50K	Total
Divorced	108 90.064	14 31.936	122
Married	246 334.42	207 118.58	453
Never-marr	276 215.56	16 76.438	292
Separated	21 15.503	0 5.4973	21
Widowed	23 18.456	2 6.5444	25
Total	674	239	913

Table of MS_new by Salary

Statistic	DF	Value	Prob
Chi-Square	4	179.3990	<.0001
Likelihood Ratio Chi-Square	4	200.2206	<.0001
Mantel-Haenszel Chi-Square	1	36.8110	<.0001
Phi Coefficient		0.4433	
Contingency Coefficient		0.4052	
Cramer's V		0.4433	

Tests for Association

Occupation

Occupation	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Adm-clerical	109	11.94	109	11.94
Craft-repair	138	15.12	247	27.05
Exec-managerial	118	12.92	365	39.98
Farming-fishing	30	3.29	395	43.26
Handlers-cleaners	45	4.93	440	48.19
Machine-op-inspct	53	5.81	493	54.00
Other-service	107	11.72	600	65.72
Priv-house-serv	5	0.55	605	66.27
Prof-specialty	123	13.47	728	79.74
Protective-serv	20	2.19	748	81.93
Sales	111	12.16	859	94.09
Tech-support	25	2.74	884	96.82
Transport-moving	29	3.18	913	100.00

Descriptive Statistics

Table of Occupation by Salary

Occupation	Salary		
	<=50K	>50K	Total
Adm-clerical	97 80.467	12 28.533	109
Craft-repair	103 101.88	35 36.125	138
Exec-managerial	63 87.111	55 30.889	118
Farming-fishing	27 22.147	3 7.8532	30
Handlers-cleaners	40 33.22	5 11.78	45
Machine-op-inspct	46 39.126	7 13.874	53
Other-service	103 78.99	4 28.01	107
Priv-house-serv	5 3.6911	0 1.3089	5
Prof-specialty	70 90.802	53 32.198	123
Protective-serv	11 14.765	9 5.2355	20
Sales	78 81.943	33 29.057	111
Tech-support	10 18.456	15 6.5444	25
Transport-moving	21 21.409	8 7.5915	29
Total	674	239	913

Table of Occupation by Salary

Statistic	DF	Value	Prob
Chi-Square	12	119.5567	<.0001
Likelihood Ratio Chi-Square	12	129.8515	<.0001
Mantel-Haenszel Chi-Square	1	9.7584	0.0018
Phi Coefficient		0.3619	
Contingency Coefficient		0.3403	
Cramer's V		0.3619	

Tests for Association

Inference: The Occupation is somewhat evenly distributed amongst the levels and has 13 levels. The likelihood ratio chi-squared test (preferred as high cell counts) concludes that there is a significant association between the target salary and Occupation variable and hence can be interpreted as a significant predictor for the target salary.

Relationship

Inference: The Relationship is dominated by three classes “Husband”, “Not-in-family” and “Own-child” and has 6 levels in total. The likelihood ratio chi-squared test (preferred as high cell counts) concludes that there is a significant association between the target salary and Relationship variable and hence can be interpreted as a significant predictor for the target salary.

Relationship	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Husband	405	44.36	405	44.36
Not-in-family	230	25.19	635	69.55
Other-relativ	28	3.07	663	72.62
Own-child	125	13.69	788	86.31
Unmarried	94	10.30	882	96.60
Wife	31	3.40	913	100.00

Descriptive Statistics

Table of Relationship by Salary

Relationship	Salary		
	<=50K	>50K	Total
Husband	213 298.98	192 106.02	405
Not-in-family	210 169.79	20 60.208	230
Other-relativ	27 20.67	1 7.3297	28
Own-child	121 92.278	4 32.722	125
Unmarried	85 69.393	9 24.607	94
Wife	18 22.885	13 8.115	31
Total	674	239	913

Table of Relationship by Salary

Statistic	DF	Value	Prob
Chi-Square	5	189.7779	<.0001
Likelihood Ratio Chi-Square	5	207.9753	<.0001
Mantel-Haenszel Chi-Square	1	73.9833	<.0001
Phi Coefficient		0.4559	
Contingency Coefficient		0.4148	
Cramer's V		0.4559	

Tests for Association

Race_new

Race_new	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Black	93	10.19	93	10.19
Eskimo-Isl	35	3.83	128	14.02
Other	11	1.20	139	15.22
White	774	84.78	913	100.00

Descriptive Statistics

Table of Race_new by Salary

Race_new	Salary <=50K	Salary >50K	Total
Black	75 68.655	18 24.345	93
Eskimo-Isl	24 25.838	11 9.1621	35
Other	10 8.1205	1 2.8795	11
White	565 571.39	209 202.61	774
Total	674	239	913

Table of Race_new by Salary

Statistic	DF	Value	Prob
Chi-Square	3	4.6740	0.1973
Likelihood Ratio Chi-Square	3	5.1885	0.1585
Mantel-Haenszel Chi-Square	1	1.8156	0.1778
Phi Coefficient		0.0716	
Contingency Coefficient		0.0714	
Cramer's V		0.0716	

Tests for Association

Inference: The Race_new is dominated by "White" class and has 4 levels. The likelihood ratio chi-squared test (preferred as high cell counts) concludes that there is no significant association between the target salary and Race_new variable and hence can be interpreted as an insignificant predictor for the target salary.

Sex

Sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	282	30.89	282	30.89
Male	631	69.11	913	100.00

Descriptive Statistics

Table of Sex by Salary

Sex	Salary <=50K	Salary >50K	Total
Female	257 208.18	25 73.82	282
Male	417 465.82	214 165.18	631
Total	674	239	913

Table of Sex by Salary

Statistic	DF	Value	Prob
Chi-Square	1	63.2817	<.0001
Likelihood Ratio Chi-Square	1	72.6409	<.0001
Continuity Adj. Chi-Square	1	61.9921	<.0001
Mantel-Haenszel Chi-Square	1	63.2124	<.0001
Phi Coefficient		0.2633	
Contingency Coefficient		0.2546	
Cramer's V		0.2633	

Tests for Association

Inference: The Sex is dominated by the "Male" class observations and has 2 levels. The likelihood ratio chi-squared test (preferred as high cell counts) concludes that there is a significant association between the target salary and MS_new variable and hence can be interpreted as a significant predictor for the target salary.

NCountry_new

NCountry_new	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Non-US	86	9.42	86	9.42
US	827	90.58	913	100.00

Descriptive Statistics

Table of NCountry_new by Salary

NCountry_new	Salary <=50K	Salary >50K	Total
Non-US	63 63.487	23 22.513	86
US	611 610.51	216 216.49	827
Total	674	239	913

Table of NCountry_new by Salary

Statistic	DF	Value	Prob
Chi-Square	1	0.0158	0.9000
Likelihood Ratio Chi-Square	1	0.0157	0.9002
Continuity Adj. Chi-Square	1	0.0000	1.0000
Mantel-Haenszel Chi-Square	1	0.0158	0.9001
Phi Coefficient		-0.0042	
Contingency Coefficient		0.0042	
Cramer's V		-0.0042	

Tests for Association

Inference: The NCountry_new is dominated by the "US" class observations and has 2 levels. The likelihood ratio chi-squared test (preferred as high cell counts) concludes that they are independent and hence can be interpreted as an insignificant predictor for the target salary.

So, the preliminary analysis emphasized the significance of the almost all the predictors except Race_new and NCountry_new. This may be speculated as a new regrouping can make it significant and hence will be investigated in the final report. Let's try to answer the question of if the insignificance is due to the grouping or if they are independent. For that we will use the original race and NCountry variable associations with salary.

Statistics for Table of Race by Salary

Statistic	DF	Value	Prob
Chi-Square	4	9.9747	0.0409
Likelihood Ratio Chi-Square	4	12.2639	0.0155
Mantel-Haenszel Chi-Square	1	0.9906	0.3196
Phi Coefficient		0.1045	
Contingency Coefficient		0.1040	
Cramer's V		0.1045	

Tests for Association of Race

Statistics for Table of NCountry by Salary

Statistic	DF	Value	Prob
Chi-Square	27	35.7151	0.1217
Likelihood Ratio Chi-Square	27	39.9914	0.0513
Mantel-Haenszel Chi-Square	1	0.8033	0.3701
Phi Coefficient		0.1978	
Contingency Coefficient		0.1940	
Cramer's V		0.1978	

WARNING: 93% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Tests for Association of NCountry

The NCountry variable is clearly insignificant in the prediction of salary. But surprisingly the Race variable from the original dataset can be concluded significant but the categories of Race variable have pure class observations which won't be able to explain much of the target variable salary and the chi-square tests won't be valid for those cell counts. Let's re-evaluate the race-new variable with regrouping "Amer-Indian-Eskimo", "Asian-Pac-Islander" and "Other" into a new group called "Other_new" to make sure the categories of the new variable have relatively high cell counts for the chi-square tests to be valid.

Table of Race_new_1 by Salary

Race_new_1	Salary <=50K	Salary >50K	Total
Black	75 68.655	18 24.345	93
Other_new	34 33.958	12 12.042	46
White	565 571.39	209 202.61	774
Total	674	239	913

Table of Race_new_1 by Salary

Statistics for Table of Race_new_1 by Salary

Statistic	DF	Value	Prob
Chi-Square	2	2.5130	0.2846
Likelihood Ratio Chi-Square	2	2.6598	0.2645
Mantel-Haenszel Chi-Square	1	2.3336	0.1266
Phi Coefficient		0.0525	
Contingency Coefficient		0.0524	
Cramer's V		0.0525	

Tests for Association

The tests show that Race_new_1 is not significant predictor for Salary. This regrouping was done to ensure there were no rows which belong to only one class so that chi-square tests are valid tests and so we proceed with the conclusion that NCountry_new and Race_new are indeed insignificant predictors for salary. Now, we look at the correlations between the continuous predictors. As shown below, the continuous predictors Age, Ednum and Hpw are not correlated with each other.

Pearson Correlation Coefficients, N = 913
Prob > |r| under H0: Rho=0

	Age	Ednum	hpw
Age	1.00000	0.06396 0.0534	0.00230 0.9448
Ednum	0.06396 0.0534	1.00000	0.17485 <.0001
hpw	0.00230 0.9448	0.17485 <.0001	1.00000

Pearson Correlation

Spearman Correlation Coefficients, N = 913
Prob > |r| under H0: Rho=0

	Age	Ednum	hpw
Age	1.00000	0.10376 0.0017	0.04199 0.2050
Ednum	0.10376 0.0017	1.00000	0.20537 <.0001
hpw	0.04199 0.2050	0.20537 <.0001	1.00000

Spearman Correlation

Based on the preliminary analysis, we can conclude that:

- Age, Ednum and Hpw are the significant continuous variables and there also no issues of multicollinearity. They can be directly used as the predictor variables in the models.
- Education_new, WC, MS_new, Occupation, Relationship and Sex are the significant categorical variables. They can be directly used as the predictor variables in the models.

Model Selection and Interpretation for Predicting Salary:

In this project, we aim to develop and evaluate predictive models to classify whether individuals earn a salary above or below USD 50,000. The target variable, Salary, is binary, indicating whether an income of an individual exceeds the amount of US$50,000. To achieve this, we employed several modeling techniques, including logistic regression, generalized linear models with a probit link, and decision trees.

Logistic regression is a widely used statistical method for binary classification problems. It estimates the probability that a given input point belongs to a certain class by fitting a logistic function to the data. This model serves as a robust baseline due to its interpretability and simplicity. The generalized linear model with a probit link function is another approach we used. The probit model, like the logistic model, is used for binary response variables but assumes a normal cumulative distribution function. This method allows us to compare its performance against logistic regression and understand any potential improvements it offers. By comparing these models, we aim to identify the most effective approach for accurately predicting the salary category. Our analysis involves evaluating model performance using metrics such as the confusion matrix, the misclassification error and ROC curve, providing a comprehensive understanding of each model's predictive capabilities.

We will use “proc logistic” on our dataset Salary_S1_6, changing our 6 categorical variables to their respective reference levels (WC, Education_new, MS_new, Occupation, Relationship, Sex), adding on our 3 numerical variables (Age, Ednum, Hpw), and utilizing stepwise selection at sle = 0.05 and sls = 0.05 thresholds. The stepwise selection process runs through 6 iterations and selects 6 variables at the end. The summary of the stepwise selection and the parameter estimates for the final iteration are depicted below. The analysis of maximum likelihood estimates output gives us the coefficients and significance of our predictors. Note that this logistic regression model predicts the probability of belonging to the class of salary with less than 50000 USD.

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	Relationship		5	1	189.7779		<.0001
2	Ednum		1	2	99.3194		<.0001
3	hpw		1	3	12.6743		0.0004
4	Age		1	4	11.4244		0.0007
5	Occupation		12	5	32.2488		0.0013
6	Sex		1	6	8.8716		0.0029

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	7.9284	1.0396	58.1590	<.0001
Occupation	Craft-repair	1	-0.2752	0.4380	0.3948	0.5298
Occupation	Exec-managerial	1	-0.7102	0.4398	2.6072	0.1064
Occupation	Farming-fishing	1	1.5698	0.7704	4.1524	0.0416
Occupation	Handlers-cleaners	1	0.4569	0.6386	0.5119	0.4743
Occupation	Machine-op-inspct	1	0.4267	0.5856	0.5310	0.4662
Occupation	Other-service	1	0.7224	0.6527	1.2250	0.2684
Occupation	Priv-house-serv	1	9.9229	591.0	0.0003	0.9866
Occupation	Prof-specialty	1	-0.7414	0.4512	2.6996	0.1004
Occupation	Protective-serv	1	-0.6104	0.6446	0.8966	0.3437
Occupation	Sales	1	-0.3101	0.4509	0.4728	0.4917
Occupation	Tech-support	1	-1.7916	0.6342	7.9806	0.0047
Occupation	Transport-moving	1	-0.2748	0.5982	0.2111	0.6459
Relationship	Not-in-family	1	1.9845	0.3210	38.2211	<.0001
Relationship	Other-relativ	1	2.6108	1.2002	4.7324	0.0296
Relationship	Own-child	1	2.0244	0.5836	12.0334	0.0005
Relationship	Unmarried	1	1.4788	0.4673	10.0163	0.0016
Relationship	Wife	1	-1.5486	0.6296	6.0502	0.0139
Sex	Male	1	-1.2126	0.4187	8.3872	0.0038
Age		1	-0.0310	0.00887	12.1799	0.0005
Ednum		1	-0.3162	0.0533	35.1937	<.0001
hpw		1	-0.0389	0.00985	15.6134	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Occupation Craft-repair vs Adm-clerical	0.759	0.322	1.792
Occupation Exec-managerial vs Adm-clerical	0.492	0.208	1.164
Occupation Farming-fishing vs Adm-clerical	4.806	1.062	21.753
Occupation Handlers-cleaners vs Adm-clerical	1.579	0.452	5.521
Occupation Machine-op-inspct vs Adm-clerical	1.532	0.486	4.829
Occupation Other-service vs Adm-clerical	2.059	0.573	7.402
Occupation Priv-house-serv vs Adm-clerical	>999.999	<0.001	>999.999
Occupation Prof-specialty vs Adm-clerical	0.476	0.197	1.154
Occupation Protective-serv vs Adm-clerical	0.543	0.154	1.921
Occupation Sales vs Adm-clerical	0.733	0.303	1.775
Occupation Tech-support vs Adm-clerical	0.167	0.048	0.578
Occupation Transport-moving vs Adm-clerical	0.760	0.235	2.454
Relationship Not-in-family vs Husband	7.276	3.878	13.649
Relationship Other-relativ vs Husband	13.610	1.295	143.038
Relationship Own-child vs Husband	7.571	2.412	23.764
Relationship Unmarried vs Husband	4.388	1.756	10.964
Relationship Wife vs Husband	0.213	0.062	0.730
Sex Male vs Female	0.297	0.131	0.676
Age	0.970	0.953	0.987
Ednum	0.729	0.657	0.809
hpw	0.962	0.943	0.981

Before going to interpret the variables, let's look at the goodness of fit tests for the logistic regression model. Since the given data is not of the event/ trial syntax, we use the Hosmer-Lemeshow goodness of fit test. The result of the Hosmer-Lemeshow goodness of fit is given below.

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
13.4805	8	0.0964

Since the p-value of the Hosmer and Lemeshow goodness of fit test is 0.0964, which is greater than 0.05. This implies that we can reject the null hypothesis and conclude that the logistic regression model fits the data well. Given that we have a logistic regression model that fits the data well, let's look at the odds ratio and coefficient interpretation of the predictor variables. As aforementioned, 6 variables were chosen: 3 categorical, and all 3 numerical variables.

- **Occupation:** It has 13 categories, and the model chose "Adm-clerical" as the reference category. Only the "Farming-fishing" and "Tech-support" categories are significant at the level 0.05 in predicting if the given observation belongs to the class with salary less than USD 50K. The same can also be inferred from the odds ratios as their confidence intervals do not contain 1. The odds of having a salary less than 50K USD for the occupation "Farming-fishing" is 4.806 times than the occupation "Adm-clerical". The odds of having a salary less than 50K USD for the occupation "Tech-support" is 0.167 times than the occupation "Adm-clerical". "Farming-fishing" occupation increases the odds of having a salary less than 50K USD while the "Tech-support" decreases the odds of having a salary less than 50K USD.

- Interpretation: Occupation "Farming-fishing" will typically result in a salary less than 50K USD, while the other occupation "Tech-support" will be more likely to have a salary greater than 50K USD.

- **Relationship:** It has 6 categories, and the model chose "Husband" as the reference category. All the categories are significant at the level 0.05 in predicting if the given observation belongs to the class with salary less than USD 50K. The same can also be inferred from the odds ratios as their confidence intervals do not contain 1. The odds of having a salary less than 50K USD for the relationship "Not-in-family" is 7.276 times than the relationship "Husband". The odds of having a salary less than 50K USD for the relationship "Other-Relatives" is 13.61 times than the relationship "Husband". The odds of having a salary less than 50K USD for the relationship "Own-child" is 7.571 times than the relationship "Husband". The odds of having a salary less than 50K USD for the relationship "Unmarried" is 4.381 times than the relationship "Husband".
  The odds of having a salary less than 50K USD for the relationship "Wife" is 0.231 times than the relationship "Husband". All relationships increase odds of having a salary less than 50K USD except for the relationship "Wife", which decreases odds of having a salary less than 50K USD.

  - Interpretation: Relationships "Not-in-family", "Other-Relatives", "Own-child" and "Unmarried" will typically result in a salary less than 50K USD, while the relationship "Wife" is more likely to have a salary greater than 50K USD.

- **Sex:** It has two categories and "Female" was the reference level. The sex "Male" is significant at the level 0.05 in predicting if the given observation belongs to the class with salary less than USD 50K. The same can also be inferred from the odds ratios as their confidence intervals do not contain 1. The odds of having a salary less than 50K USD for the sex "Male" is 0.297 times than the sex "Female".

  - Interpretation: Females are more likely to have a salary less than 50K USD than males according to this dataset.

- **Age:** This continuous variable is significant at the level 0.05 in predicting if the given observation belongs to the class with salary less than USD 50K. The odds of having a salary less than 50K USD increases 0.97 times with a unit increase in age.

  - Interpretation: As age increases, the salaries will more likely be greater than 50K according to this dataset.

- **Ednum:** This continuous variable is significant at the level 0.05 in predicting if the given observation belongs to the class with salary less than USD 50K. The odds of having a salary less than 50K USD increases 0.729 times with a unit increase in Ednum.

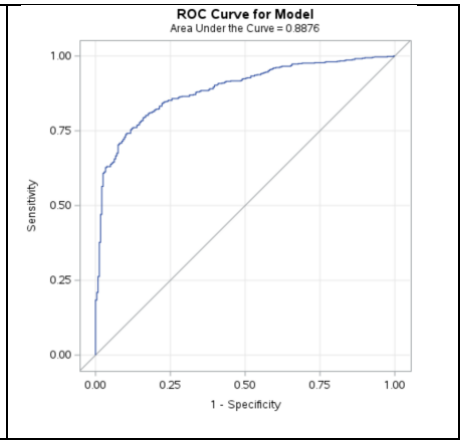
  - Interpretation: As Ednum increases, the salaries will more likely be greater than 50K according to this dataset.

- **<u>Hpw:</u>** This continuous variable is significant at the level 0.05 in predicting if the given observation belongs to the class with salary less than USD 50K. The odds of having a salary less than 50K USD increases 0.962 times with a unit increase in Hpw.

    - <u>Interpretation:</u> As Hpw increases, the salaries will more likely be greater than 50K according to this dataset.

Below given are the model evaluation results of the logistic regression model.

Table of Salary by _INTO_			
	INTO(Formatted Value of the Predicted Response)		
Salary	<=50K	>50K	Total
<=50K	612	62	674
>50K	101	138	239
Total	713	200	913



The generalized linear model framework, as implemented in the GENMOD procedure in SAS, extends traditional linear modeling to accommodate response variables that follow various distributions, including the normal, binomial, Poisson, and others. This flexibility allows the modeling of data that exhibit non-normal behavior, which is common in real-world scenarios. One of the key features of GENMOD is its ability to use different link functions to relate the mean of the response variable to the linear predictor. We observed that using the logit link function and the binomial distribution, we were getting the same results as the logistic regression model fitted using proc logistic. Then, we utilized the probit link function, which is particularly useful for binary outcome variables. The probit link models the probability of the outcome occurring as a function of the normal cumulative distribution function (CDF). This approach assumes that the latent variable underlying the binary outcome follows a normal distribution, providing an alternative to the commonly used logit link function. The probit model can offer better fit and interpretation in cases where the data exhibit characteristics more aligned with the normal distribution. By using the GENMOD procedure with the probit link, we aimed to improve the predictive accuracy and interpretability of our model, particularly in comparison to logistic regression. The generalized linear model was fitted with probit link and all the significant predictors from the initial exploratory analysis. The results showed slight improvements in model performance metrics such as the ROC curve, indicating that the probit link was a suitable choice for our binary salary prediction task.

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	6.9609	2.2986	2.4558	11.4660	9.17	0.0025
WC	Local-gov	1	0.4533	0.3645	-0.2611	1.1677	1.55	0.2136
WC	Private	1	0.2778	0.3019	-0.3139	0.8695	0.85	0.3575
WC	Self-emp-inc	1	0.1491	0.4130	-0.6604	0.9586	0.13	0.7181
WC	Self-emp-not-inc	1	0.6423	0.3459	-0.0357	1.3203	3.45	0.0633
WC	State-gov	1	0.4915	0.4156	-0.3230	1.3060	1.40	0.2370
Education_new	Bachelors	1	0.4307	0.3669	-0.2884	1.1498	1.38	0.2404
Education_new	Doctorate	1	0.2205	1.0372	-1.8124	2.2533	0.05	0.8317
Education_new	HS-grad	1	-0.5793	0.5418	-1.6412	0.4826	1.14	0.2849
Education_new	Masters	1	0.1226	0.5481	-0.9517	1.1969	0.05	0.8230
Education_new	Prof-schoo	1	0.4786	0.7746	-1.0395	1.9968	0.38	0.5366
Education_new	School	1	-0.9825	1.0544	-3.0490	1.0841	0.87	0.3514
Education_new	Some-colle	1	-0.4980	0.3926	-1.2676	0.2715	1.61	0.2046
MS_new	Married	1	-0.3774	0.5420	-1.4398	0.6849	0.48	0.4862
MS_new	Never-marr	1	-0.0234	0.2668	-0.5464	0.4996	0.01	0.9300
MS_new	Separated	1	5.4299	7609.730	-14909.4	14920.23	0.00	0.9994
MS_new	Widowed	1	-0.2576	0.5264	-1.2894	0.7741	0.24	0.6245
Occupation	Craft-repair	1	-0.2163	0.2641	-0.7339	0.3014	0.67	0.4129
Occupation	Exec-managerial	1	-0.5332	0.2680	-1.0586	-0.0078	3.96	0.0467
Occupation	Farming-fishing	1	0.5335	0.4159	-0.2817	1.3487	1.65	0.1996
Occupation	Handlers-cleaners	1	0.2970	0.3704	-0.4290	1.0229	0.64	0.4227
Occupation	Machine-op-inspct	1	0.2162	0.3481	-0.4660	0.8984	0.39	0.5345
Occupation	Other-service	1	0.3611	0.3645	-0.3534	1.0756	0.98	0.3220
Occupation	Priv-house-serv	1	4.3561	15510.11	-30394.9	30403.62	0.00	0.9998
Occupation	Prof-specialty	1	-0.5332	0.2732	-1.0687	0.0023	3.81	0.0510
Occupation	Protective-serv	1	-0.5740	0.4109	-1.3793	0.2313	1.95	0.1624
Occupation	Sales	1	-0.2890	0.2727	-0.8234	0.2455	1.12	0.2893
Occupation	Tech-support	1	-1.1798	0.3763	-1.9173	-0.4423	9.83	0.0017
Occupation	Transport-moving	1	-0.1987	0.3618	-0.9078	0.5105	0.30	0.5829
Relationship	Not-in-family	1	0.7033	0.5635	-0.4011	1.8077	1.56	0.2120
Relationship	Other-relativ	1	0.8315	0.6994	-0.5393	2.2022	1.41	0.2345
Relationship	Own-child	1	0.7460	0.6244	-0.4778	1.9698	1.43	0.2322
Relationship	Unmarried	1	0.5117	0.5423	-0.5513	1.5746	0.89	0.3454
Relationship	Wife	1	-0.9842	0.3628	-1.6952	-0.2732	7.36	0.0067
Sex	Male	1	-0.7592	0.2391	-1.2279	-0.2905	10.08	0.0015
Age		1	-0.0175	0.0054	-0.0280	-0.0070	10.67	0.0011
Ednum		1	-0.3696	0.1853	-0.7328	-0.0064	3.98	0.0461
hpw		1	-0.0212	0.0056	-0.0322	-0.0103	14.38	0.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

Out of these predictors, only occupation "Exec-managerial", occupation "Tech-support", relationship "Wife", sex "Male", age, Ednum and Hpw are significant in predicting the probability of having the salary less than 50K USD. The reference categories chosen were "federal-government" for WC, "Associate" for Education_new, "Divorced" for MS_new, "Adm-clerical" for occupation, "Husband" for relationship, and "Female" for sex.

Before going to interpret the variables, let's look at the goodness of fit tests for the generalized linear model. Since the given data is not of the event/ trial syntax, we use the Hosmer-Lemeshow goodness of fit test. The result of the Hosmer-Lemeshow goodness of fit is given below.

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
14.1394	8	0.0782

Since the p-value of the Hosmer and Lemeshow goodness of fit test is 0.0782, which is greater than 0.05. This implies that we can reject the null hypothesis and conclude that the generalized linear model fits the data well. Given that we have a generalized linear model that fits the data well, let's look at the coefficient interpretation of the predictor variables. As aforementioned, 6 variables were chosen: 6 categorical, and all 3 numerical variables.

- **WC:** It has 6 categories, and the model chose "Federal-government" as the reference category. None of the categories are significant at the level 0.05 in predicting if the given observation belongs to the class with salary less than USD 50K. All the coefficients are positive indicating that they increase the likelihood of the event that it belongs to the class with salary less than USD 50K.

    - Interpretation: WC may not be able to explain if the salary is less or greater than USD 50K as none of the categories are significant. Overall, all the categories almost indicate that they increase the likelihood of the event that it belongs to the class with salary less than USD 50K.

- **Education_new:** It has 8 categories, and the model chose "Associate" as the reference category. None of the categories are significant at the level 0.05 in predicting if the given observation belongs to the class with salary less than USD 50K. The coefficients of the categories "Bachelors", "Doctorate", "Masters", and "Professional school" are positive indicating that they increase the likelihood of the event that it belongs to the class with salary less than USD 50K. The rest of the coefficients are negative indicating that they decrease the likelihood of the event that it belongs to the class with salary less than USD 50K.

    - Interpretation: Education_new may not be able to explain if the salary is less or greater than USD 50K as none of the categories are significant. Overall, some of the categories indicate that they increase the likelihood of the event that it belongs to the class with salary less than USD 50K while the others indicate they decrease the likelihood of the event that it belongs to the class with salary less than USD 50K.

- **MS_new:** It has 5 categories, and the model chose "Divorced" as the reference category. None of the categories are significant at the level 0.05 in predicting if the given observation belongs to the class with salary less than USD 50K. All the coefficients are negative except "Separated" indicating that they decrease the likelihood of the event that it belongs to the class with salary less than USD 50K.

    - Interpretation: MS_new may not be able to explain if the salary is less or greater than USD 50K as none of the categories are significant. Overall, all the categories except "Separated" indicate that they decrease the likelihood of the event that it belongs to the class with salary less than USD 50K while the "Separated" category increase the likelihood of the event that it belongs to the class with salary less than USD 50K.

- **Occupation:** It has 13 categories, and the model chose "Adm-clerical" as the reference category. Only the "Exec-managerial" and "Tech-support" categories are significant at the level 0.05 in predicting if the given observation belongs to the class with salary less than USD 50K. The coefficients of both "Exec-managerial" and "Tech-support" categories are negative indicating that they decrease the likelihood of the event that it belongs to the class with salary less than USD 50K.

    - Interpretation: Occupations "Exec-managerial" and "Tech-support" categories will typically result in a salary greater than 50K USD.

- **Relationship:** It has 6 categories, and the model chose "Husband" as the reference category. Only the "Wife" category is significant at the level 0.05 in predicting if the given observation belongs to the class with salary less than USD 50K. The coefficient of the "Wife" category is negative indicating that it decreases the likelihood of the event that it belongs to the class with salary less than USD 50K.

    - Interpretation: Relationship "Wife" typically result in a salary greater than 50K USD.

- **Sex:** It has two categories and "Female" was the reference level. The sex "Male" is significant at the level 0.05 in predicting if the given observation belongs to the class with salary less than USD 50K. The coefficient of the "Male" category is negative indicating that it decreases the likelihood of the event that it belongs to the class with salary less than USD 50K.

    - Interpretation: Males are more likely to have a salary greater than 50K USD than females according to this dataset.

- **Age:** This continuous variable is significant at the level 0.05 in predicting if the given observation belongs to the class with salary less than USD 50K. The coefficient of age is negative indicating that it decreases the likelihood of the event that it belongs to the class with salary less than USD 50K.

    - Interpretation: As age increases, the salaries will more likely be greater than 50K according to this dataset.

  Ednum: This continuous variable is significant at the level 0.05 in predicting if the given observation belongs to the class with salary less than USD 50K. The coefficient of Ednum is negative indicating that it decreases the likelihood of the event that it belongs to the class with salary less than USD 50K.

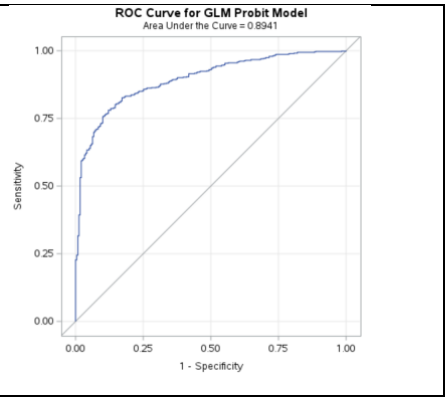
    - Interpretation: As Ednum increases, the salaries will more likely be greater than 50K according to this dataset.

Hpw: This continuous variable is significant at the level 0.05 in predicting if the given observation belongs to the class with salary less than USD 50K. The coefficient of Hpw is negative indicating that it decreases the likelihood of the event that it belongs to the class with salary less than USD 50K.

- Interpretation: As Hpw increases, the salaries will more likely be greater than 50K according to this dataset.

Below given are the model evaluation results of the generalized linear model.

Table of Salary by pred_class			
	pred_class		
Salary	<=50K	>=50K	Total
<=50K	607	67	674
>50K	92	147	239
Total	699	214	913



Comparing the results of both the logistic regression and generalized linear model are given below.

Metric	Model	
	Logistic Regression	GLM with Probit Link
<= 50 K Misclassifications	62	67
>= 50 K Misclassifications	101	92
Misclassification Error	17.85%	17.41%
F1Score	62.86%	64.85%
AUC	88.76%	89.41%

The above table suggests that the generalized linear model with probit link performs marginally better than the logistic regression model. But this improved performance comes at the cost of the ease of interpretation. So, considering the ease of feature selection and interpretation in logistic regression, we will be using the logistic regression for inferences and conclusions.

Conclusions

The logistic regression model with stepwise selection has selected six predictors: Occupation, Relationship, Sex, Age, Ednum and Hpw. The validation of their significance was observed

through the p-values of the parameter estimates of the generalized linear model with probit link on the full model with all the predictors. Although occupation is significant but only two of its categories “Farming-fishing” and “Tech-support” are significant at the level 0.05 in predicting if the given observation belongs to the class with salary less than USD 50K. The conclusions are that occupation “Farming-fishing” will result in a salary less than 50K USD as the odds of having a salary less than 50K USD for the occupation “Farming-fishing” is 4.806 times than the occupation “Adm-clerical”, while the other occupation “Tech-support” will be more likely to have a salary greater than 50K USD as the odds of having a salary greater than 50K USD for the occupation “Tech-support” is 5.988 times than the occupation “Adm-clerical”. The conclusions are that relationships “Not-in-family”, “Other-Relatives”, “Own-child” and “Unmarried” will result in a salary less than 50K USD as the odds of having a salary less than 50K USD are at least 4 times than the relationship “Husband”, while the relationship “Wife” will result in a salary greater than 50K USD as the odds of having a salary greater than 50K USD for the relationship “Wife” is 4.694 times than the relationship “Husband”. The conclusions based on sex are that females are more likely to have a salary less than 50K USD compared to males as the odds of having a salary less than 50K USD for the sex “Female” is 3.367 times than the sex “Male”. For a unit increase in age, the salaries will more likely be greater than 50K as the odds of having a salary greater than 50K USD increases 1.03 times according to this dataset. For a unit increase in number of years of education, the salaries will more likely be greater than 50K as the odds of having a salary greater than 50K USD increases 1.37 times according to this dataset. For a unit increase in hours per week, the salaries will more likely be greater than 50K as the odds of having a salary greater than 50K USD increases 1.04 times according to this dataset.

In summary, the logistic regression model with stepwise selection has identified key predictors of salary, providing valuable insights into the factors that influence whether an individual's salary is above or below 50K USD. The significance of predictors such as occupation, relationship status, sex, age, education level, and hours worked per week underscores the complex interplay of demographic and socio-economic variables in determining income levels. The analysis highlights specific occupational categories, relationship statuses, and demographic factors that significantly impact salary expectations. These findings can inform targeted policy interventions, career counseling, and individual career planning to address income disparities and enhance economic outcomes for various demographic groups. By understanding the significant predictors and their effects on salary, stakeholders can develop strategies to support higher income potential, particularly for groups identified as having higher odds of earning less than 50K USD.

Contributions:

Both of us contributed equally to the modeling and exploratory analysis of the project. We collaboratively worked on data preprocessing, exploring key trends and relationships in the dataset. In the modeling phase, we both participated in selecting and fitting the appropriate models, including logistic regression with stepwise selection, as well as interpreting the significance of the predictors. Our joint efforts were integral in validating the models and analyzing the results, ensuring that the final report accurately reflected our findings and insights.