OPIM 5604: Predictive Modeling

White Paper

Team 7

Anand Tamhankar | Shan Xu | Shruthi Tadamiri | Urvi Kohli

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY

Our project is based on a scenario in which a technology company, PineApple Inc., has a marketing budget of $100,000 to be spent a social media campaign to sell its $1,000 laptop. The primary aim of the project is to identify key variables from the data available on social media websites to target customers who can afford the product. Our target customer is a person having an income greater than $50,000.

For our project, we used the census data as a proxy for social media profile information. The original data set was pretty large, containing 32,561 samples, making it easier for us to find meaningful patterns and trends. As the data came from the census bureau, it was of good quality and reliable. Only about 5% of the rows had missing values, which we removed. Also, most of the variables were important and not correlated, making them useful for our models.

We ran seven models on the data set: logistic regression, decision tree, bootstrap forest, boosted tree, neural network, K-Neighborhood Network, and Naïve Bayes. All the models had a similar misclassification rate between 16 – 17%. However, for the sake of simplicity and distinct ability to make suggestions in a business scenario, we chose the decision tree model to explore further and check for trends.

The optimal decision tree model was represented by a massive tree structure with 60 splits that yielded a misclassification rate of 16.73%. However, for the sake of decision making, we pruned the tree further down to 5 splits, which yielded a misclassification rate of 18.18%. Therefore, we made a conscious decision to sacrifice least amount of accuracy for greater simplicity and have provided our recommendations using the pruned tree that contains 5 splits.

2. BUSINESS PROBLEM

The key business problem is allocation of budget for the marketing campaign. We have considered a scenario in which we are advising a technology company, PineApple Inc. to sell their laptops. So, the business case is that the company is undertaking a marketing campaign of $100,000 on social media website.

The laptops are to be sold at $1,000 per unit. As the ticket price of the item is substantially high, all the people would not have income to afford the product. Hence, our aim is to target people who have a high income. Using the census and income data, our target customers are those that have annual income greater than $50,000. We used this measure as a proxy for a person who would be able to buy the product.

Further, we know that the cost to reach one customer is $1. This means that at the end of the marketing campaign, we are going to reach 100,000 people. Put in simpler terms, the business problem at hand is the limited budget of PineApple Inc., which forces it to reach only 100,000 people with its campaign. How can they maximize the impact of their campaign using the limited budget by choosing the correct variables to target high income customers?

3. SEMMA OVERVIEW

The dataset we chose for our project is “Adult census income”. This data was extracted from the Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics).

The purpose of the analysis is to predict if a person has an income of more than $50,000 a year. The result of the analysis is then used for a social media campaign that targets the people with an income higher than $50,000, who are more likely to purchase a laptop worth $1000.

3.1 SAMPLE

The population consisted of census data for all individuals and details about their income. A sample of this data was chosen for our analysis using the following conditions:

1. Age of individual is between 16 to 100
2. Number of hours worked is greater than 0

Sample size: 32,561,

Number of variables: 15

The table below lists the variables in the data set and their data types:

Variable Name	Description
Age	Continuous
Workclass	Categorical with levels: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
Fnlwgt	Continuous
Education	Categorical with levels: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
education-num	Continuous
marital-status	Categorical with levels: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
Occupation	Categorical with levels: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
Relationship	Categorical with levels: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
race	Categorical with levels: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
sex	Categorical with levels: Female, Male
capital-gain	Continuous
capital-loss	Continuous
hours-per-week	Continuous
native-country	Categorical with levels: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos,

	Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands
income	Categorical with levels - <=50k, >50k

No further sampling was done on this data set with 32,561 records.

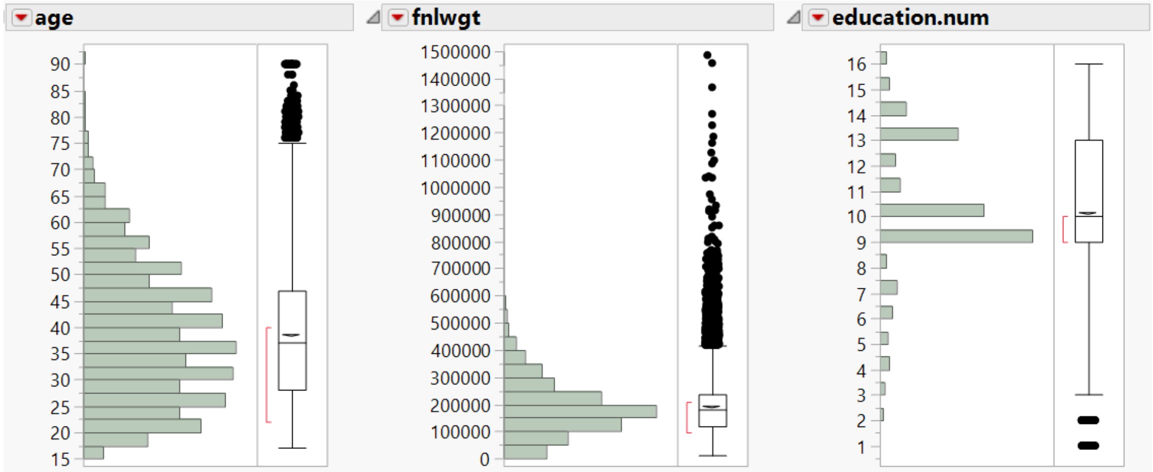
3.2 EXPLORE

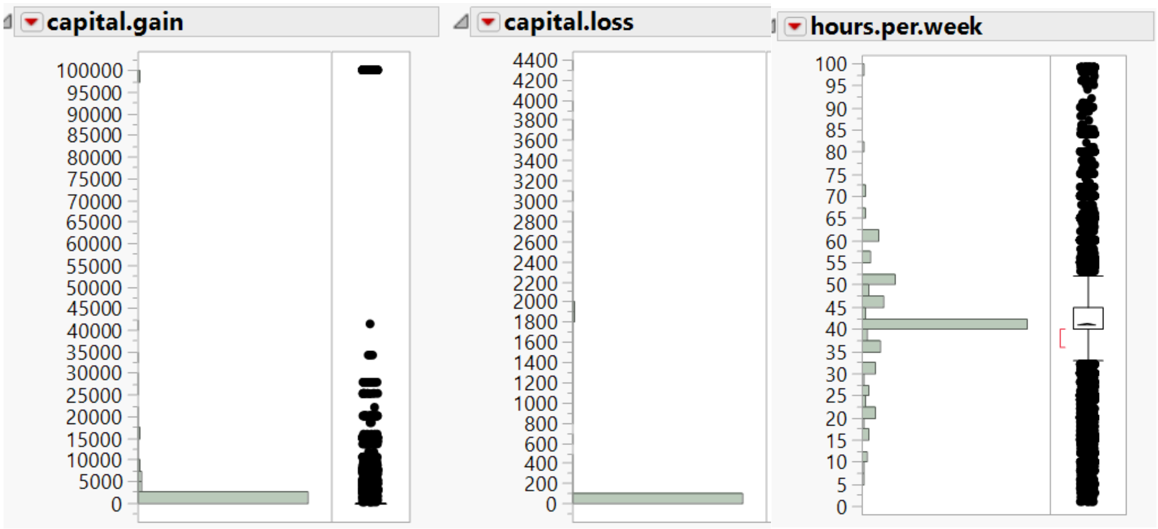
The outcome variable was identified as income, which is categorical and has two levels: <=$50k and >$50k.

We performed various exploration methods to get a sense of the data. These methods have been described below briefly:

a) OUTLIER ANALYSIS

There were only 6 numerical variables in our data set and hence we performed an outlier analysis on these 6 variables. The results are represented in the screenshots below:

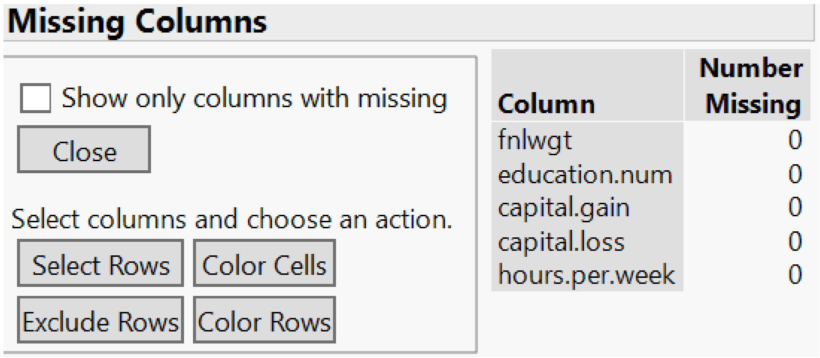




As we can see, the variables – age, fnlwgt and capital gain and hours per week indicated high number of outliers.

a) MISSING VALUES

The numerical variables were checked for missing values and it was seen that they did not contain any.



Missing Columns

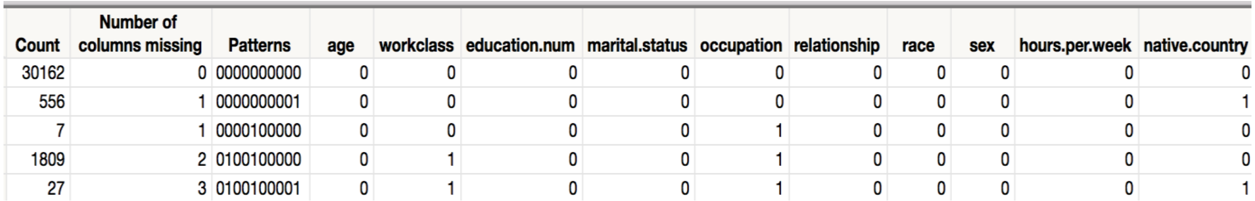
□ Show only columns with missing
Close
Select columns and choose an action.
Select Rows
Color Cells
Exclude Rows
Color Rows

Column	Number Missing
fnlwgt	0
education.num	0
capital.gain	0
capital.loss	0
hours.per.week	0

However, the following categorical variables had 2399 missing values in total:

- workingclass
- occupation
- native country

Missing data pattern:

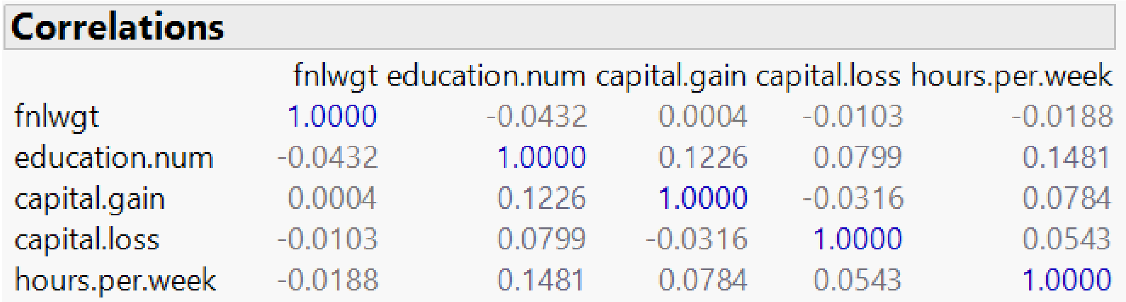


Count	Number of columns missing	Patterns	age	workclass	education.num	marital.status	occupation	relationship	race	sex	hours.per.week	native.country
30162	0	0000000000	0	0	0	0	0	0	0	0	0	0
556	1	0000000001	0	0	0	0	0	0	0	0	0	1
7	1	0000100000	0	0	0	0	1	0	0	0	0	0
1809	2	0100100000	0	1	0	0	1	0	0	0	0	0
27	3	0100100001	0	1	0	0	1	0	0	0	0	1

This constituted only 7.36% of the entire data.

b) CORRELATION

The numerical variables were analyzed for correlation and the screenshot below shows the correlation matrix for these variables:



Correlations

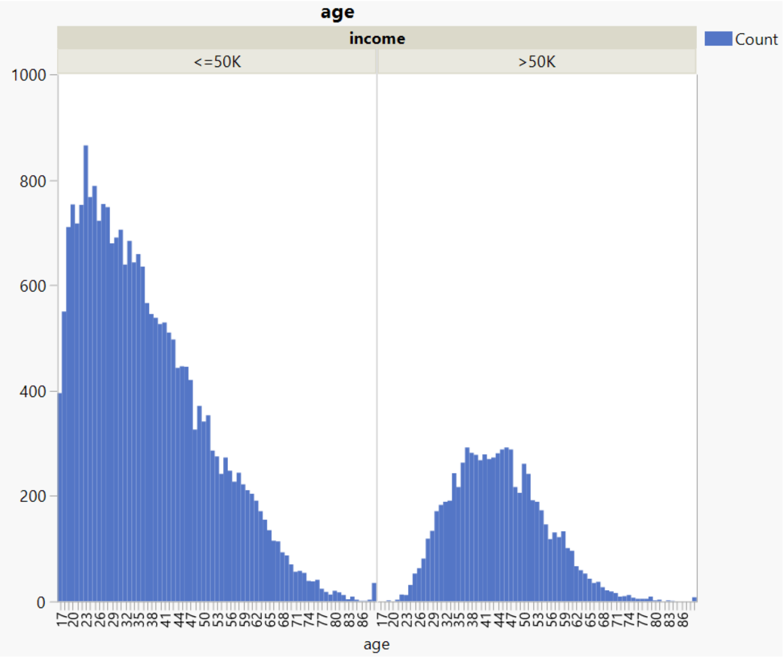
	fnlwgt	education.num	capital.gain	capital.loss	hours.per.week
fnlwgt	1.0000	-0.0432	0.0004	-0.0103	-0.0188
education.num	-0.0432	1.0000	0.1226	0.0799	0.1481
capital.gain	0.0004	0.1226	1.0000	-0.0316	0.0784
capital.loss	-0.0103	0.0799	-0.0316	1.0000	0.0543
hours.per.week	-0.0188	0.1481	0.0784	0.0543	1.0000

As we can see, none of the variables had any significant correlation between them.

c) GENERAL VISUALIZATIONS

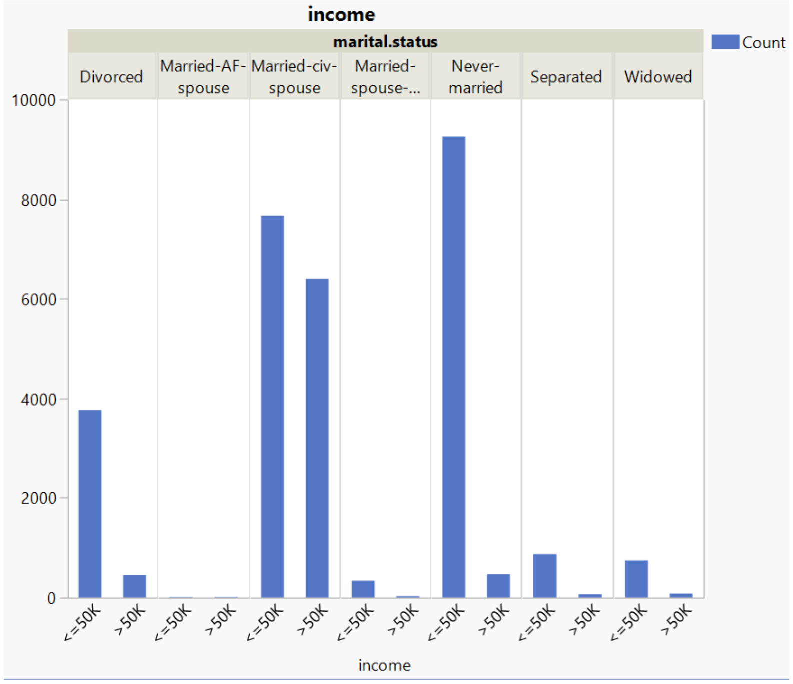
The relationships between the outcome variable – income, and the different predictors was assessed.

Income vs age



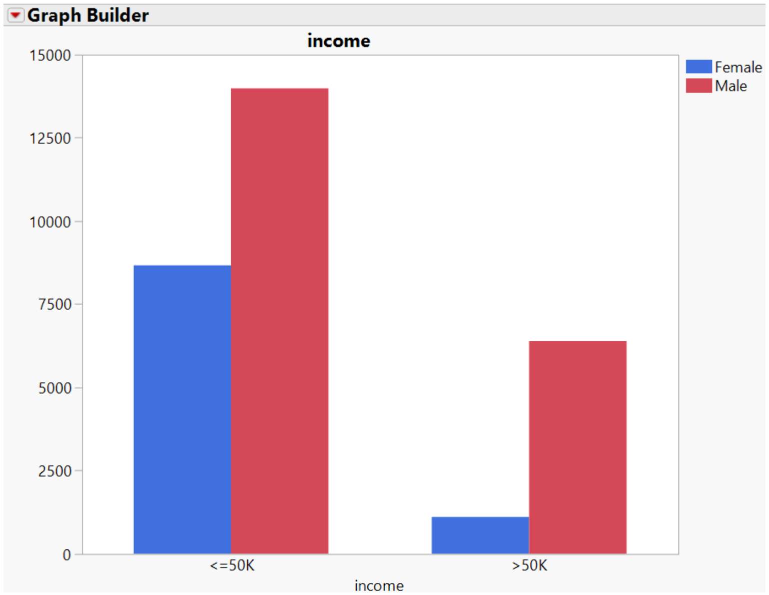
It can be inferred that there was a majority in the number of people with lower ages who had an income less than $50k. While, the number of people who had an income greater than $50k were relatively older.

Income vs marital status



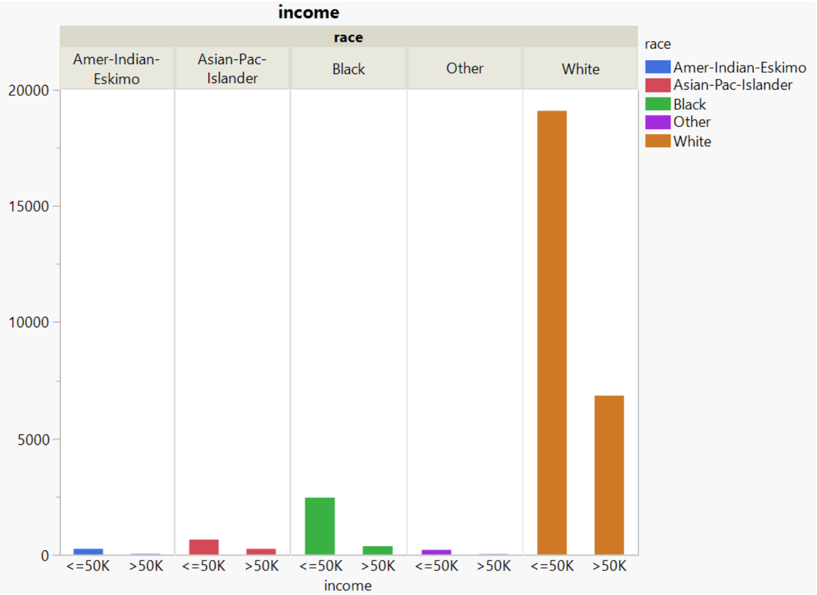
It was observed that there were more number of married people having an income greater than $50k. People who were never married were more likely to have an income less than $50k.

Income vs sex



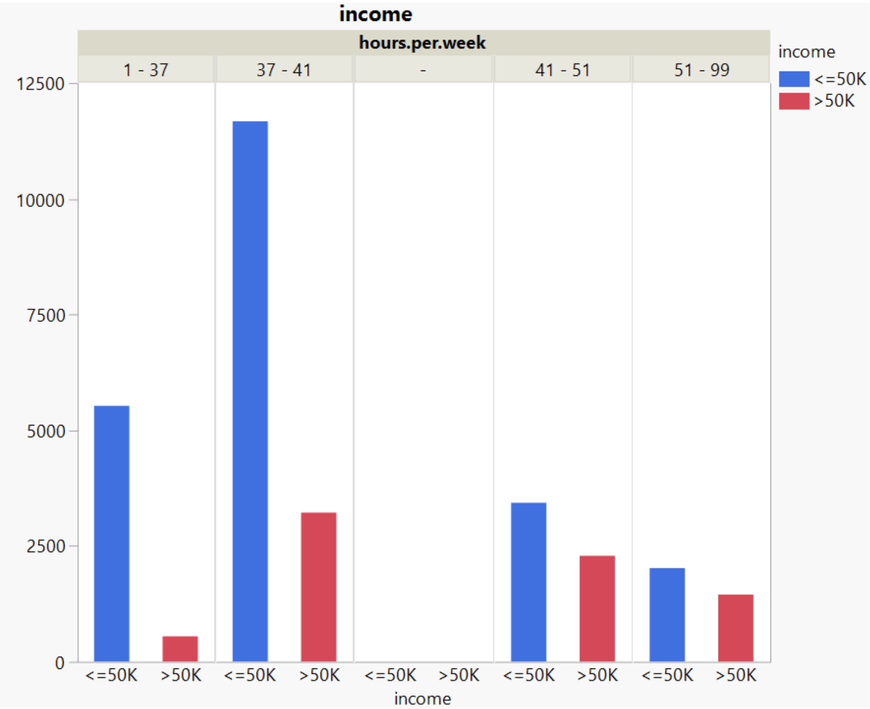
It is clear that men earn significantly higher than women in both income brackets. There are more men who earn less $50k than less than $50k.

Income vs race



A majority of the white people had income in both brackets. All other races were significantly lower.

Income vs hours per week



People who work between 37-41 hours per week have a high possibility of earning more than $50k.

d) VARIABLES REDUCTION

For the marketing campaign, we were looking for data that is easily available on an individual's social media profiles. The variables fnlwgt, capital.gain, and capital.loss were parameters that could not be extracted from social media profiles. For this reason, we dropped these three variables from our analysis for a more efficient model.

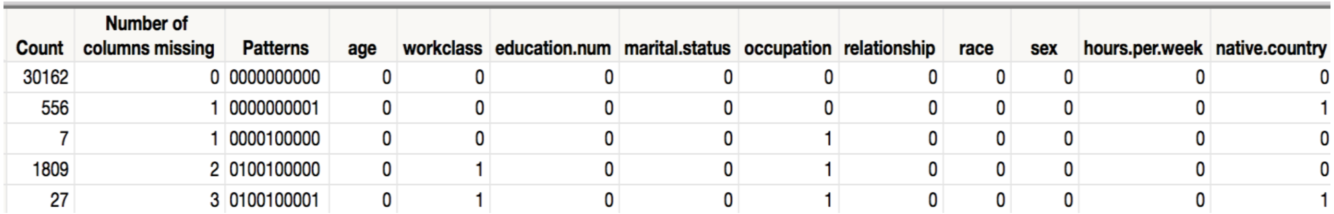
The categorical variable education was directly related to the numerical variable – education.num. This variable simply converted the various levels of education into numbers. Therefore, the education variable was dropped to avoid issues of correlation.

3.3 MODIFY

MISSING VALUE TREATMENT

In our dataset, all missing values were from categorical variables. The missing values represented only 7.36% of the entire dataset. As this was a fairly low percentage, we simply deleted all the missing values.

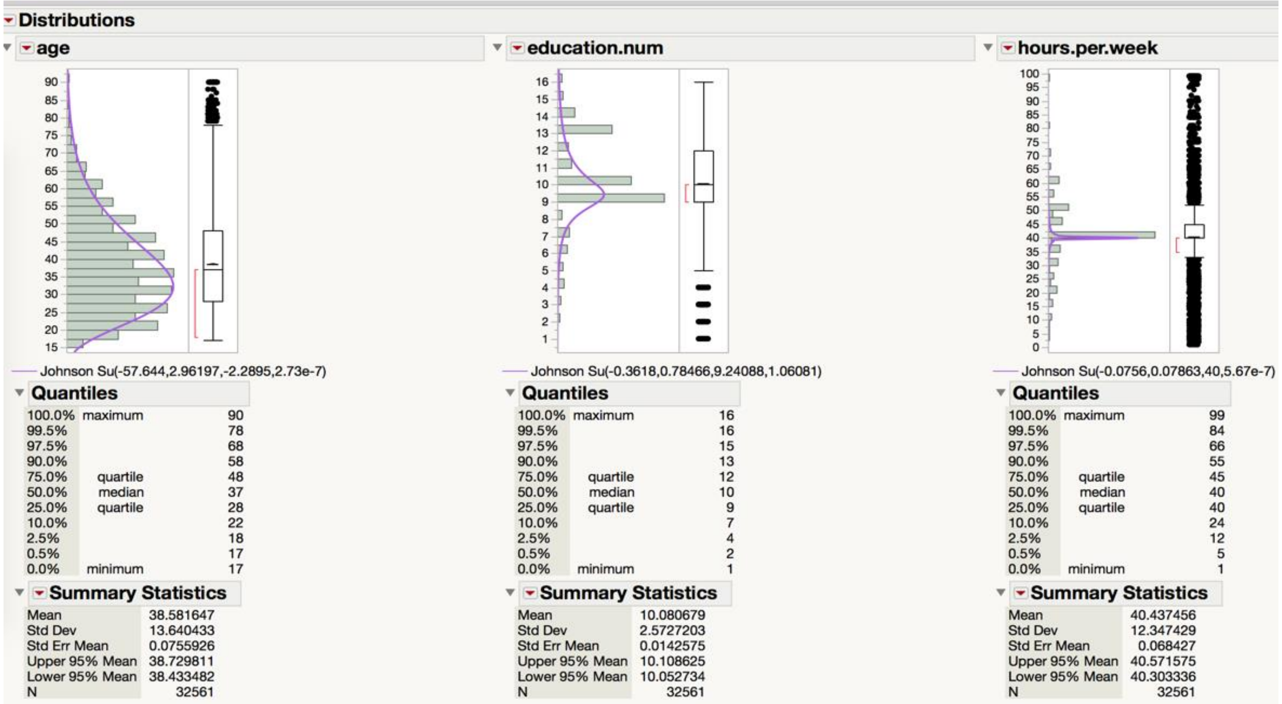
The screenshot below represents the missing data pattern in the dataset:



Count	Number of columns missing	Patterns	age	workclass	education.num	marital.status	occupation	relationship	race	sex	hours.per.week	native.country
30162	0	0000000000	0	0	0	0	0	0	0	0	0	0
556	1	0000000001	0	0	0	0	0	0	0	0	0	1
7	1	0000100000	0	0	0	0	1	0	0	0	0	0
1809	2	0100100000	0	1	0	0	1	0	0	0	0	0
27	3	0100100001	0	1	0	0	1	0	0	0	0	1

OUTLIER TREATMENT

To check for the presence of outliers, we looked at only the 6 numerical variables. Out of the 6 numerical variables, 3 had outliers, as shown in the screenshot below. We transformed these 3 numerical variables, and the screenshot of variables after transformation is shown below. As we notice from the screenshot below, the outliers of the transformed variables did not change much after transformation. Also, only hours.per.week had outliers that represented a very small percentage (2.2%) in the entire dataset. Therefore, we ignored these outliers as they were insignificant in impacting our large dataset.



Distributions

age

Johnson Su(-57.644,2.96197,-2.2895,2.73e-7)

Quantiles

100.0%	maximum	90
99.5%		78
97.5%		68
90.0%		58
75.0%	quartile	48
50.0%	median	37
25.0%	quartile	28
10.0%		22
2.5%		18
0.5%		17
0.0%	minimum	17

Summary Statistics

Mean	38.581647
Std Dev	13.640433
Std Err Mean	0.0755926
Upper 95% Mean	38.729811
Lower 95% Mean	38.433482
N	32561

education.num

Johnson Su(-0.3618,0.78466,9.24088,1.06081)

Quantiles

100.0%	maximum	16
99.5%		16
97.5%		15
90.0%		13
75.0%	quartile	12
50.0%	median	10
25.0%	quartile	9
10.0%		7
2.5%		4
0.5%		2
0.0%	minimum	1

Summary Statistics

Mean	10.080679
Std Dev	2.5727203
Std Err Mean	0.0142575
Upper 95% Mean	10.108625
Lower 95% Mean	10.052734
N	32561

hours.per.week

Johnson Su(-0.0756,0.07863,40,5.67e-7)

Quantiles

100.0%	maximum	99
99.5%		84
97.5%		66
90.0%		55
75.0%	quartile	45
50.0%	median	40
25.0%	quartile	40
10.0%		24
2.5%		12
0.5%		5
0.0%	minimum	1

Summary Statistics

Mean	40.437456
Std Dev	12.347429
Std Err Mean	0.068427
Upper 95% Mean	40.571575
Lower 95% Mean	40.303336
N	32561

OTHER MODIFICATIONS TO THE DATASET

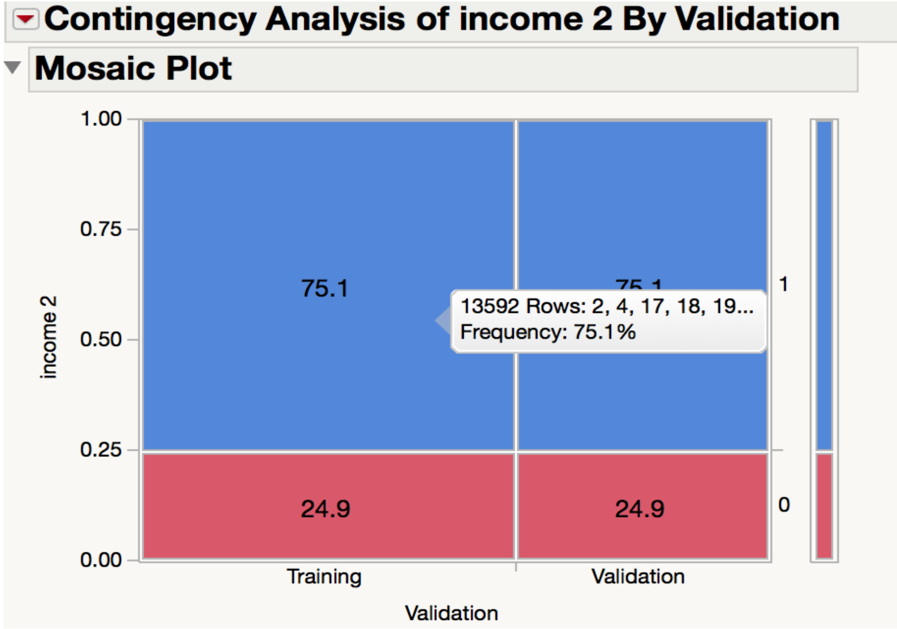
To make the analysis easier, we binned the Age variable into ranges of 5, such as 22-26, 27-31 and so on. Additionally, we dropped the education variable as it represents the same information as education.num and including it in the set of variables would lead to multicollinearity. Lastly, fnlwgt, capital.gain, and capital.loss were variables representing data that cannot be easily extracted from social media profiles. Therefore, we did not see any point in including these variables in our analysis. In other words, they are not considered as factors that contribute to or help our model in making predictions.

3.4 MODEL

<u>Splitting the dataset into Training and Validation sets</u>

We created a validation column using Stratified Sampling on the target variable. For splitting the data set into training and validation sets, we used a 60%:40% ratio.

Since our target variable income is a categorical variable with two levels (>50k and <50k), both the training and validation sets have equal proportion of 1s and 0s of the target variable. The results can be seen in the screenshot of the mosaic plot below.



Contingency Analysis of income 2 By Validation

Mosaic Plot

13592 Rows: 2, 4, 17, 18, 19...
Frequency: 75.1%

Model Building

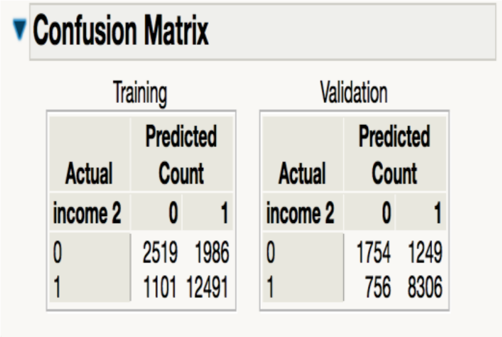
For the purpose of analysis, we ran the following seven models on the data set:

- Logistic Regression
- Neural Network
- Decision Tree
- Boosted Tree
- Bootstrap Forest
- K-Nearest Neighbor
- Naïve Bayes

The misclassification rate of the validation dataset for all seven models used in analysis are shown in the table below.

Model	Logistic Regression	Neural Network	Decision Tree	Boosted Tree	Bootstrap Forest	KNN	Naïve Bayes
Missclassification rate	0.1662	0.1653	0.1673	0.1690	0.1615	0.1728	0.2027

The screenshots below show the confusion matrix of all seven models.



Confusion Matrix

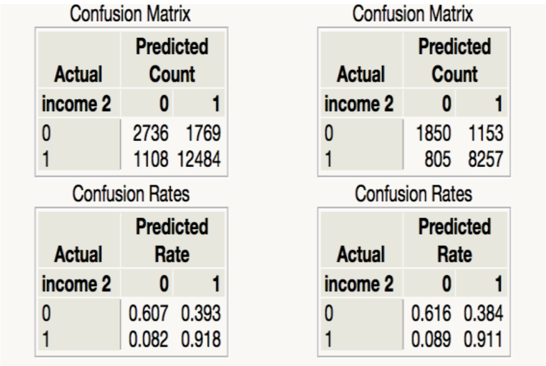
Training

Actual	Predicted Count	
income 2	0	1
0	2519	1986
1	1101	12491

Validation

Actual	Predicted Count	
income 2	0	1
0	1754	1249
1	756	8306

Logistic Regression



Confusion Matrix

Actual	Predicted Count	
income 2	0	1
0	2736	1769
1	1108	12484

Confusion Rates

Actual	Predicted Rate	
income 2	0	1
0	0.607	0.393
1	0.082	0.918

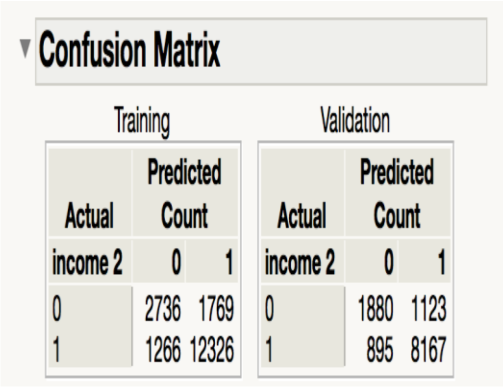
Confusion Matrix

Actual	Predicted Count	
income 2	0	1
0	1850	1153
1	805	8257

Confusion Rates

Actual	Predicted Rate	
income 2	0	1
0	0.616	0.384
1	0.089	0.911

Neural Network



Confusion Matrix

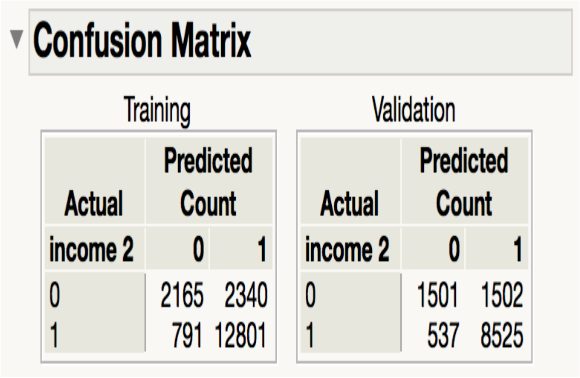
Training

Actual	Predicted Count	
income 2	0	1
0	2736	1769
1	1266	12326

Validation

Actual	Predicted Count	
income 2	0	1
0	1880	1123
1	895	8167

Decision Tree



Confusion Matrix

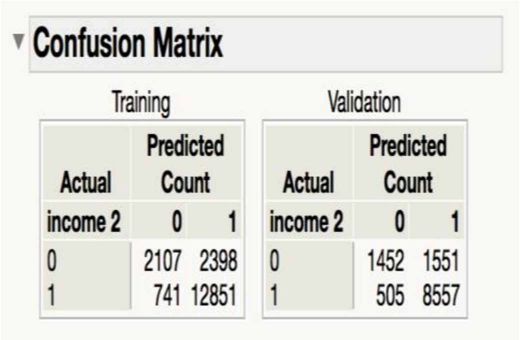
Training

Actual	Predicted Count	
income 2	0	1
0	2165	2340
1	791	12801

Validation

Actual	Predicted Count	
income 2	0	1
0	1501	1502
1	537	8525

Boosted Tree



Confusion Matrix

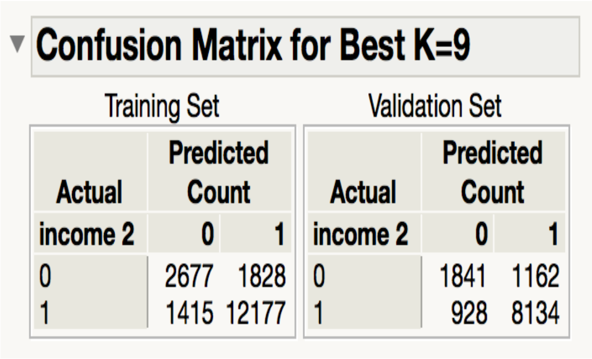
Training

Actual	Predicted Count	
income 2	0	1
0	2107	2398
1	741	12851

Validation

Actual	Predicted Count	
income 2	0	1
0	1452	1551
1	505	8557

Boosted Forest



Confusion Matrix for Best K=9

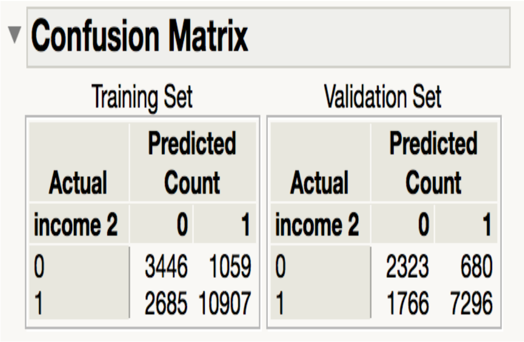
Training Set

Actual	Predicted Count	
income 2	0	1
0	2677	1828
1	1415	12177

Validation Set

Actual	Predicted Count	
income 2	0	1
0	1841	1162
1	928	8134

K-Nearest Neighbor



Confusion Matrix

Training Set

Actual	Predicted Count	
income 2	0	1
0	3446	1059
1	2685	10907

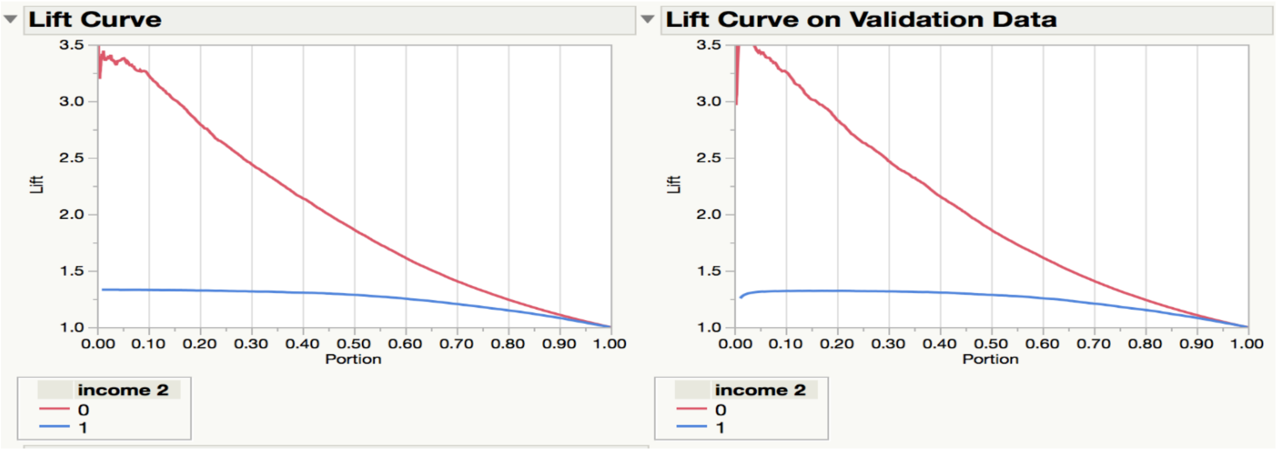
Validation Set

Actual	Predicted Count	
income 2	0	1
0	2323	680
1	1766	7296

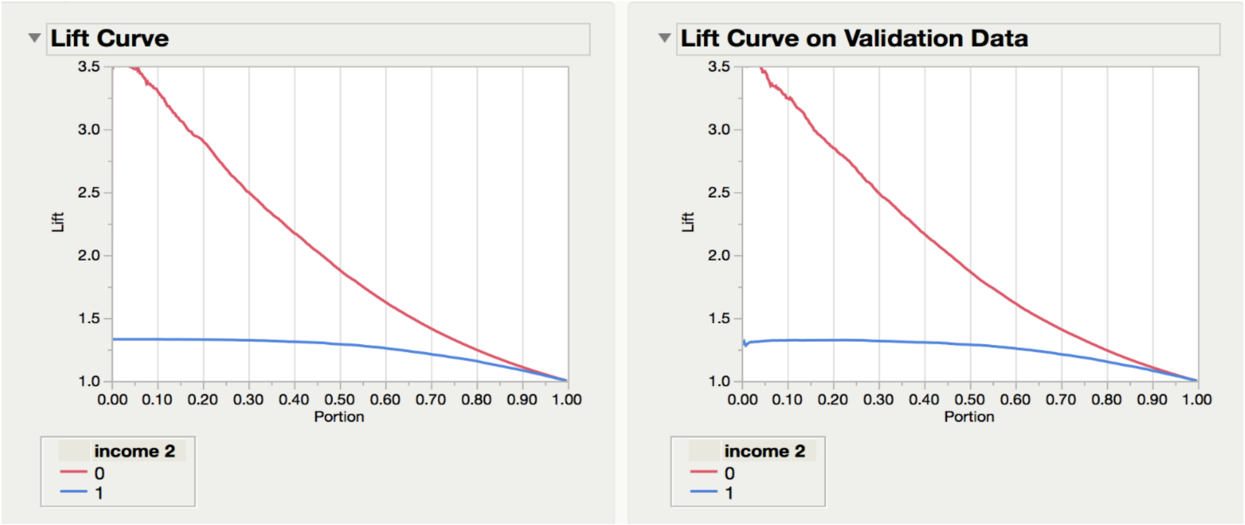
Naïve Bayes

We also analyzed Lift Curve to find out the most important class compared to the baseline model.

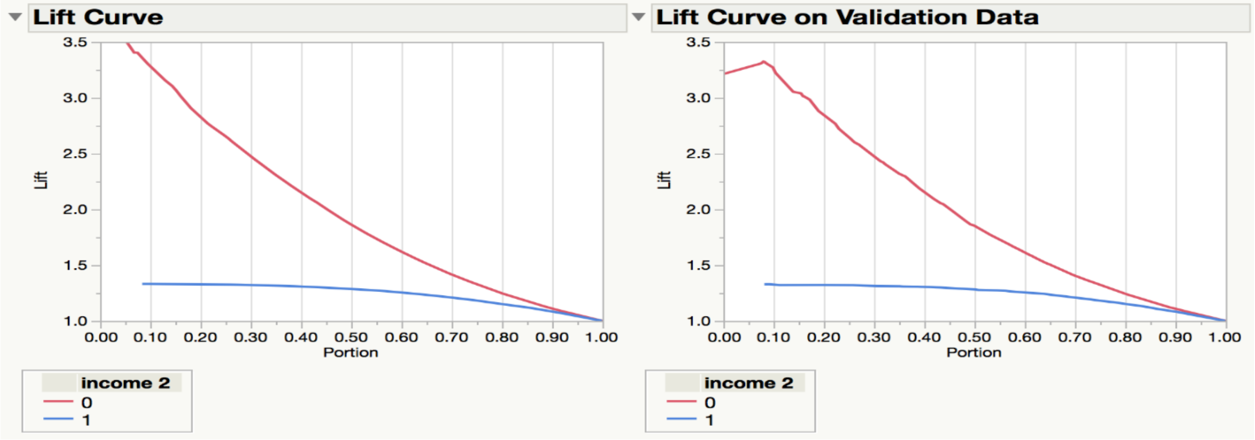
The screenshots below represent the Lift Curves for the models.



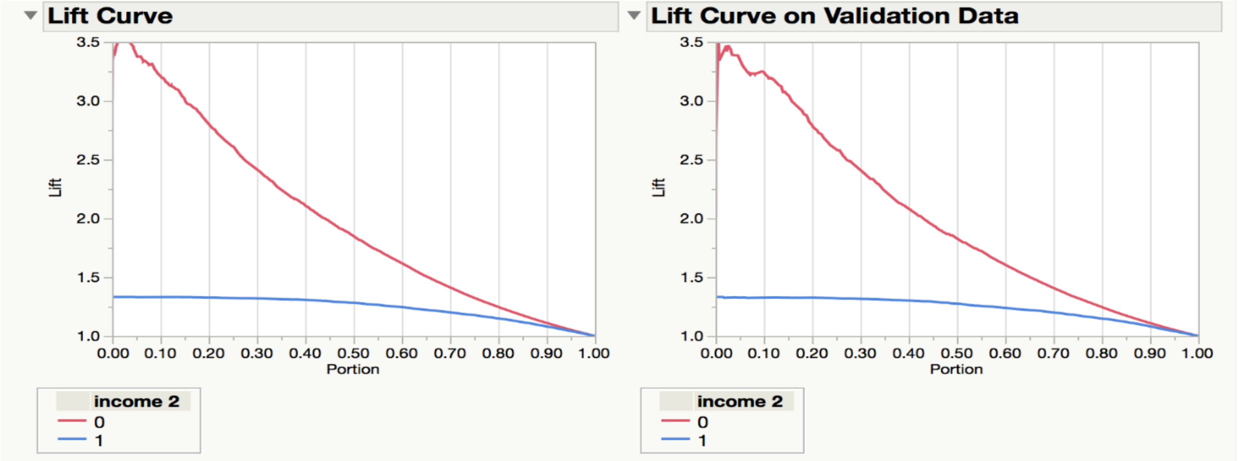
Logistic Regression



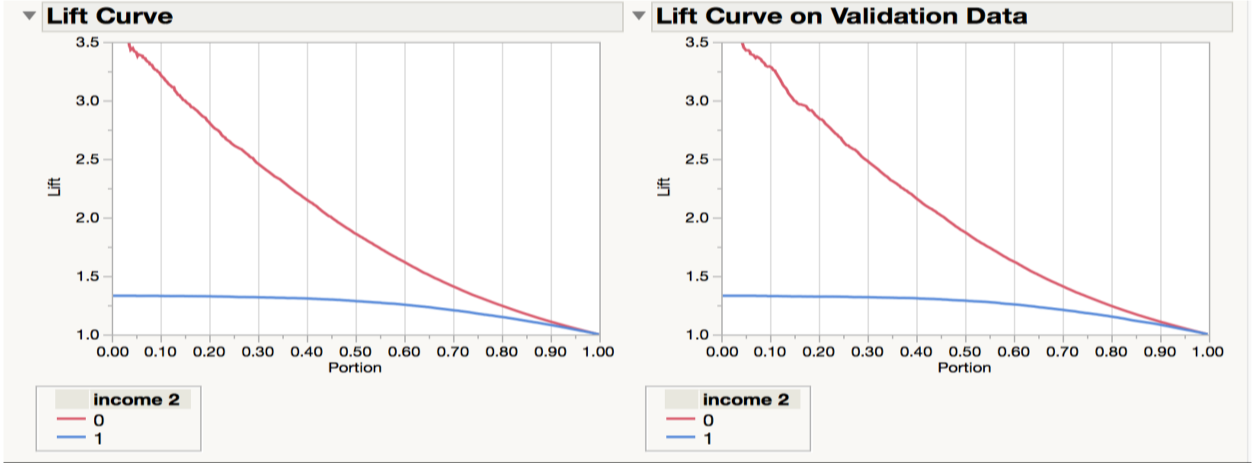
Neural Network



Decision Tree



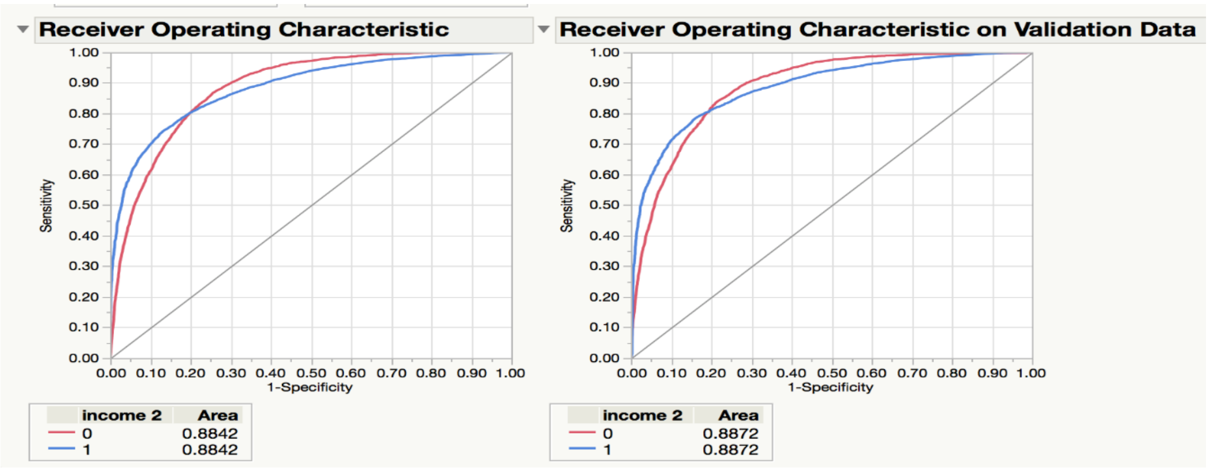
Bootstrap Forest



Boosted tree

Furthermore, we analyzed the ROC Curves to evaluate the strength of our models. All the ROC curves are shown below.

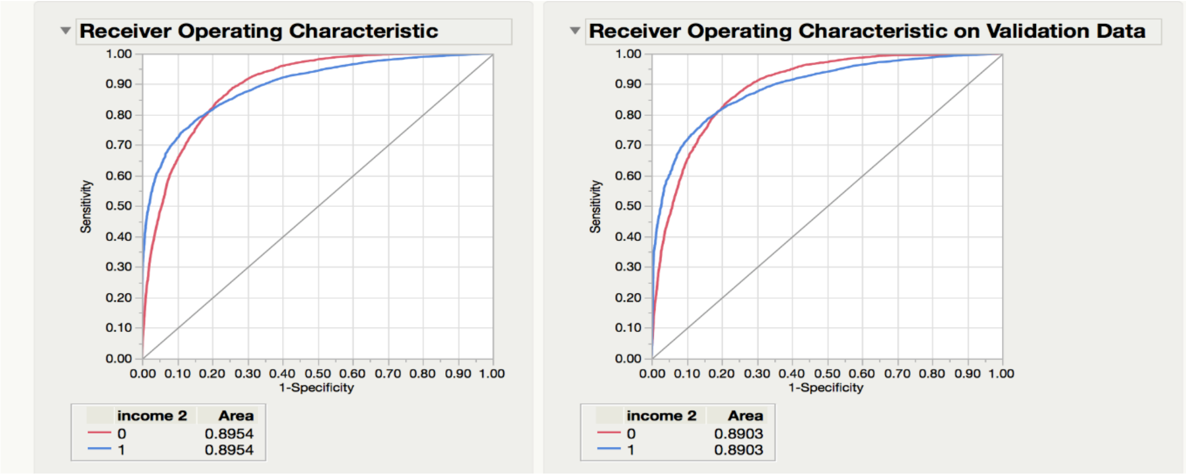
From the rule of ROC curves, we know that if ROC Index > 0.7, the model is considered to be a strong model. The models had ROC Index of approx. 0.88, suggesting that all models were strong enough.



income 2	Area
🟥 0	0.8842
🟦 1	0.8842

income 2	Area
🟥 0	0.8872
🟦 1	0.8872

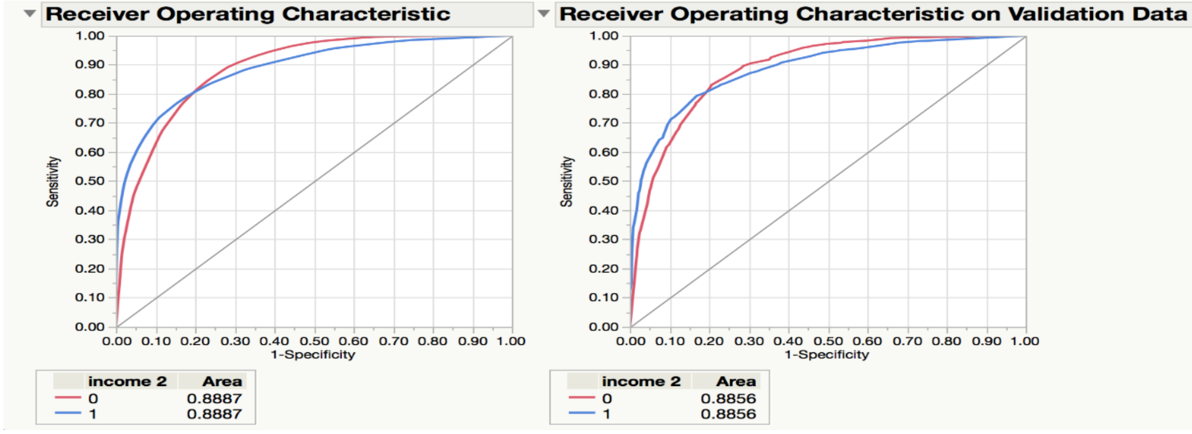
Logistic Regression



income 2	Area
🟥 0	0.8954
🟦 1	0.8954

income 2	Area
🟥 0	0.8903
🟦 1	0.8903

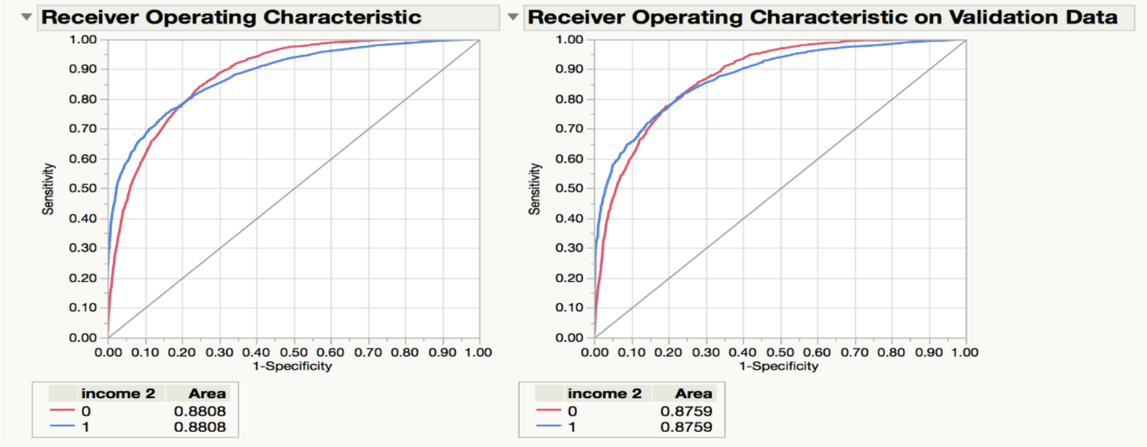
Neural Network



income 2	Area
🟥 0	0.8887
🟦 1	0.8887

income 2	Area
🟥 0	0.8856
🟦 1	0.8856

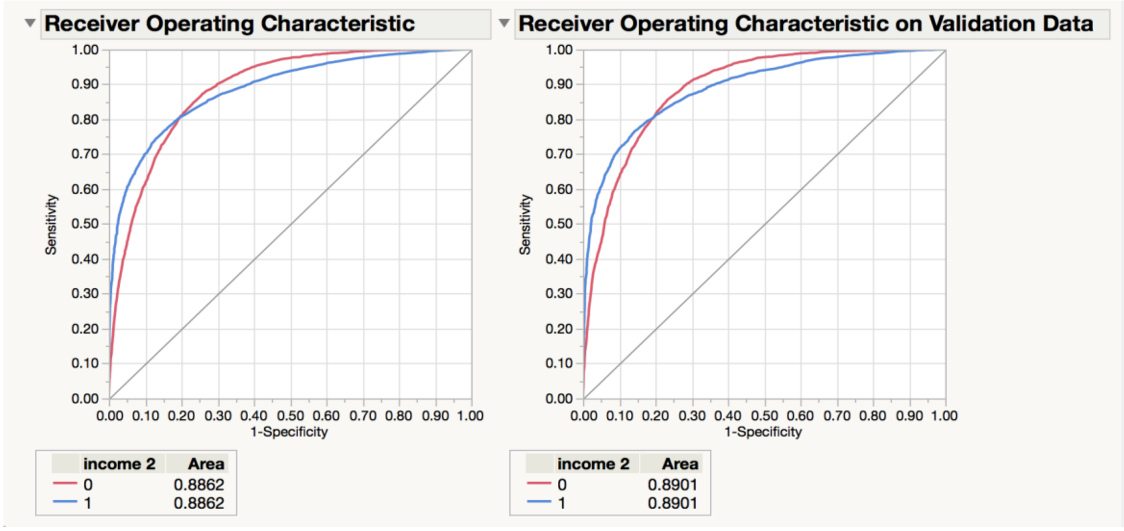
Decision tree



income 2	Area
0	0.8808
1	0.8808

income 2	Area
0	0.8759
1	0.8759

Bootstrap Forest



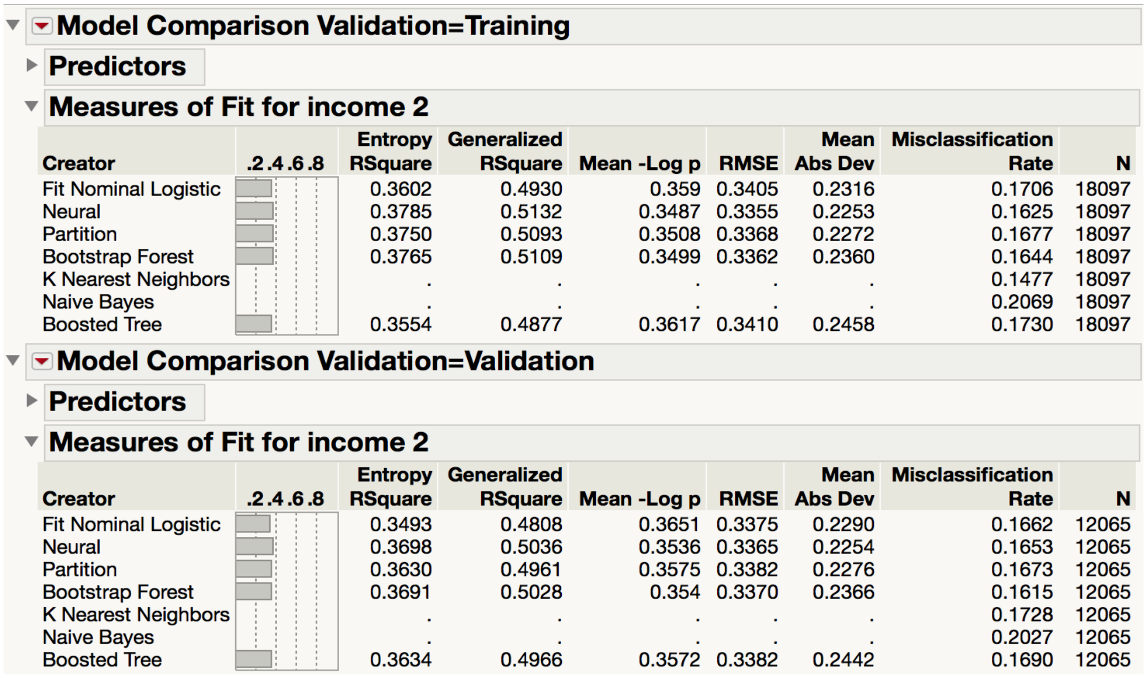
income 2	Area
0	0.8862
1	0.8862

income 2	Area
0	0.8901
1	0.8901

Boosted Tree

3.5 ASSESS

When we assessed the seven chosen models using model comparison, we saw a small variation in the values of training and validation data sets. This indicates that there was no overfitting of data. Moreover, we found that the neural and partition (decision tree) models had the lowest misclassification rate and the highest R-Square. However, as we know from our knowledge of data mining, Neural Networks are considered as “black box” prediction machines, and they cannot provide any insight into the relationships between the outcome and prediction variables. Therefore, we chose Logistic Regression and Decision Tree as our final models.



Model Comparison Validation=Training

Predictors

Measures of Fit for income 2

Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N
Fit Nominal Logistic		0.3602	0.4930	0.359	0.3405	0.2316	0.1706	18097
Neural		0.3785	0.5132	0.3487	0.3355	0.2253	0.1625	18097
Partition		0.3750	0.5093	0.3508	0.3368	0.2272	0.1677	18097
Bootstrap Forest		0.3765	0.5109	0.3499	0.3362	0.2360	0.1644	18097
K Nearest Neighbors		.	.	.	.	.	0.1477	18097
Naive Bayes		.	.	.	.	.	0.2069	18097
Boosted Tree		0.3554	0.4877	0.3617	0.3410	0.2458	0.1730	18097

Model Comparison Validation=Validation

Predictors

Measures of Fit for income 2

Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N
Fit Nominal Logistic		0.3493	0.4808	0.3651	0.3375	0.2290	0.1662	12065
Neural		0.3698	0.5036	0.3536	0.3365	0.2254	0.1653	12065
Partition		0.3630	0.4961	0.3575	0.3382	0.2276	0.1673	12065
Bootstrap Forest		0.3691	0.5028	0.354	0.3370	0.2366	0.1615	12065
K Nearest Neighbors		.	.	.	.	.	0.1728	12065
Naive Bayes		.	.	.	.	.	0.2027	12065
Boosted Tree		0.3634	0.4966	0.3572	0.3382	0.2442	0.1690	12065

4. BUSINESS VALUE

To assess the business value of any strategy, it has to be compared to a situation in which we can analyze what would happen if we face the situation without any strategy. If the cost to design any strategy is higher than the possible benefits if can reap, then the whole strategy is of no use, as it does not create any value for business.

In the business case we chose for this project, the primary challenge for PineApple Inc. is the limited budget. The company can only spend $100,000 on the marketing campaign. As we know, it takes $1 to reach a customer, therefore, at the end of marketing campaign, the company can reach only 100,000 people.

The business' ultimate motive is to increase the sales which are directly proportional to targeting the right set of people through the marketing campaign.

First, let us assume that the company goes into a marketing campaign without any plan. From the census data, we know that only 24.9% population has an annual income above $50,000. Hence, in such a situation the company's marketing campaign can only reach 24,900 people who would most likely buy the product. In other words, more than 75% of the company's budget is wasted. Now, let us assume that only 1% customers who read the advertisement can buy the product so the expected revenue in this scenario would be $249,000. Thus, expenditure of $100,000 would generate a sale of $249,000.

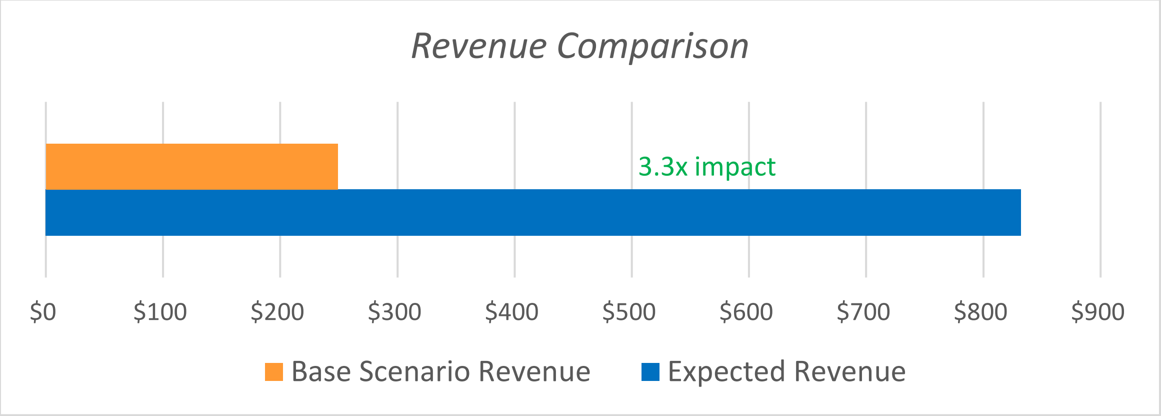
Moving on, if we consider the confusion matrix from our tree model as shown below:

Actual Count	**Predicted Count**		
Income	**0**	**1**	**Total**
0	1261	1742	**3003**
1	452	8610	**9062**
Total	**1713**	**10352**	**12065**

Using the confusion matrix above, we notice that from the 10,352 people predicted to have income >50k, it is likely that 8,610 people will have this income (i.e. >50k). Thus, the new identification rate is 83.2%.

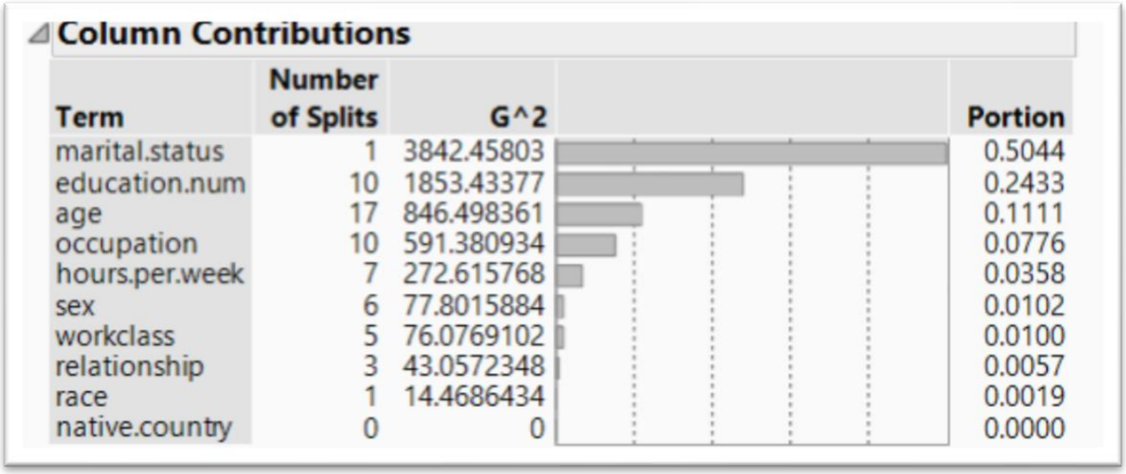
Let us see how that translates to business value. In the new model, the identification rate is 83.2%. Therefore, out of the 100,000 people targeted, we will correctly reach 83,200 people who have an income over $50,000. Let us again assume a conversion rate of 1%. Therefore, in this case, the estimated sales would be $832,000.

If we compare the sales using this new model ($832,000) to the baseline model ($249,000) ("no strategy" model), we figure that the $100,000 marketing campaign is 3.3 times more impactful than the baseline model. The revenue different is shown in the screenshot below.



5. RECOMMENDATIONS

Prior to the modeling of the data, the major contributors with respect to input variables were checked and the results indicated that marital status, education.num, and age were major contributors.



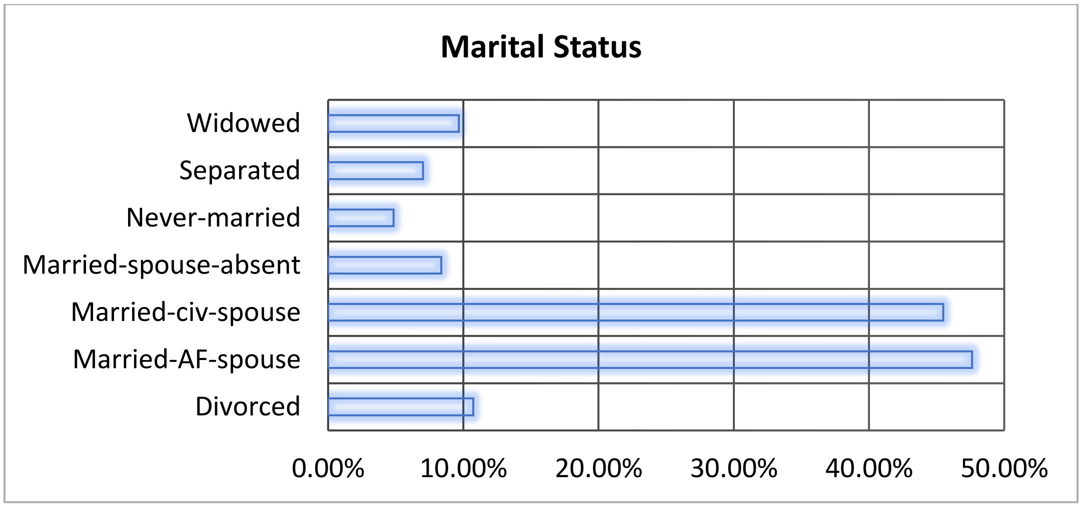
Column Contributions

Term	Number of Splits	G^2	Portion
marital.status	1	3842.45803	0.5044
education.num	10	1853.43377	0.2433
age	17	846.498361	0.1111
occupation	10	591.380934	0.0776
hours.per.week	7	272.615768	0.0358
sex	6	77.8015884	0.0102
workclass	5	76.0769102	0.0100
relationship	3	43.0572348	0.0057
race	1	14.4686434	0.0019
native.country	0	0	0.0000

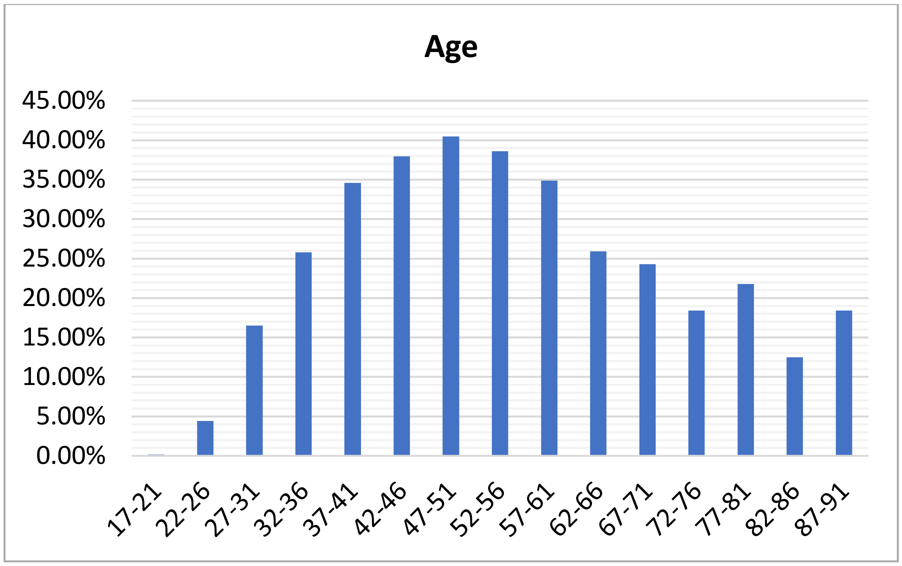
It can be inferred from the observations that the most common predictor variables were marital status, education level, occupation, age, and education level.

Further assessment of these four predictor variables revealed the following:

1. Married people had more than 40% chance of earning more than $50k per year as indicated by the statistics below.

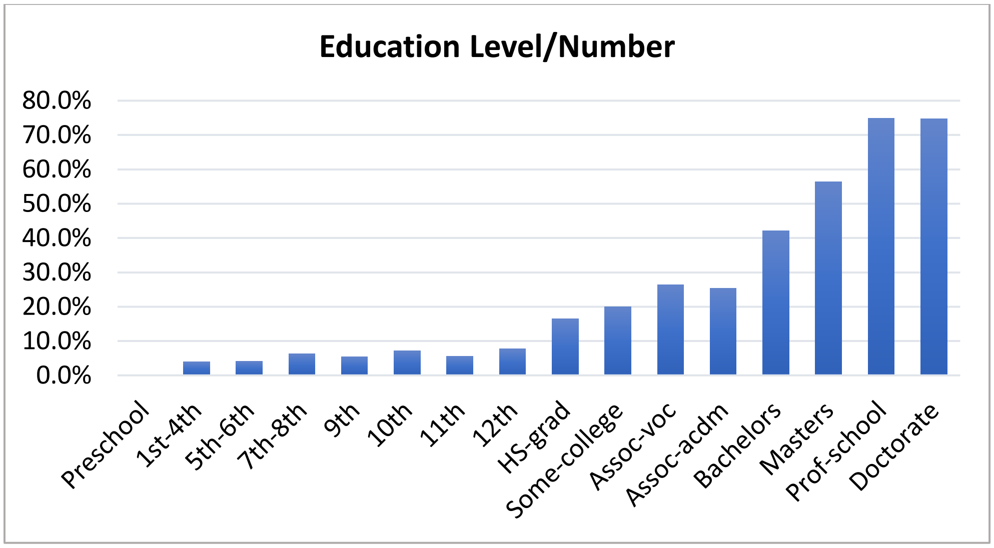


2. The age group between 47-51 had a 40% chance of earning more than $50k per year. Also, people aged between 37-41 as well as 52-56 had more than 30% chance of earning more than $50k per year.



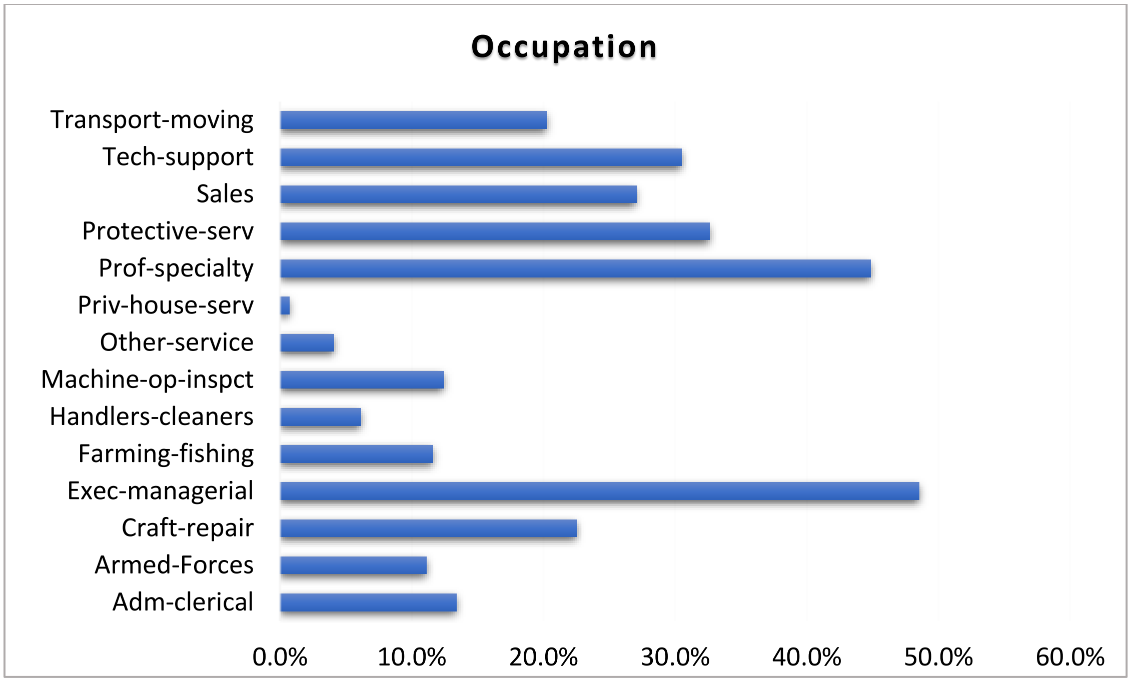
3. Education Number/Level

The more educated an individual is, higher are the chances of earning more than $50k. People who have a doctorate or have a degree from a professional school have more than 70% probability of earning more than 50k.

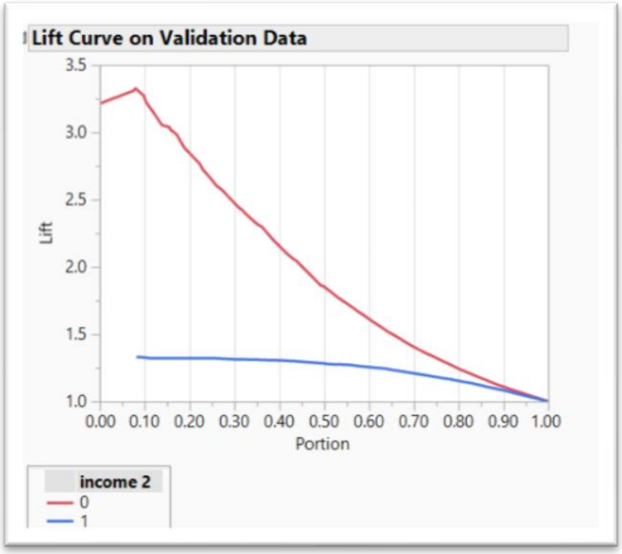


4. Occupation

People in occupations such as tech-support, sales, protective services, professional specialty services,and managerial/executive roles have a higher likelihood of having an income over $50k.



The Lift Curve for Decision Tree Model is as shown below:



According to this curve, we can conclude that the top 40% of the data is twice as likely to contain people with an annual income greater than $50k as compared to the same proportion of data with any model deployed and using random selection.

Based on the full partition tree model which was chosen, the leaf report was extracted. Specifically, the rules for the condition – earns more than $50k, were analyzed. The following observations were made:

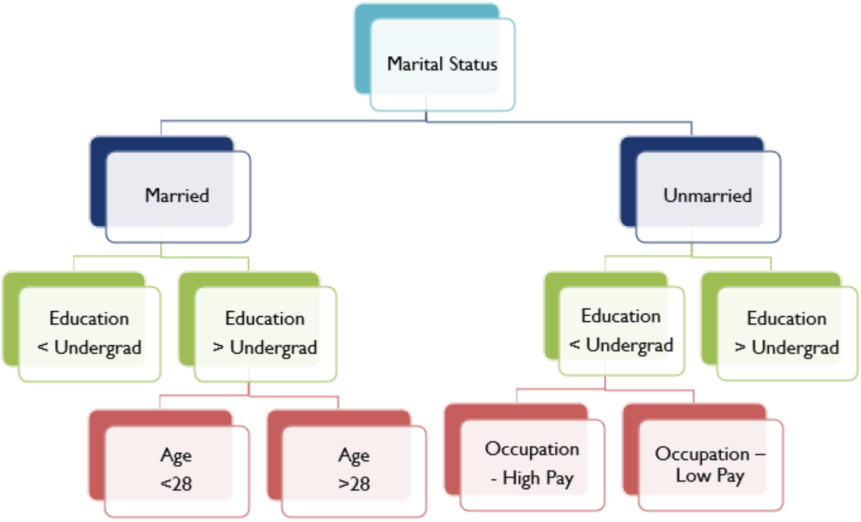
1. People who have a marital status as Married, have an education level less than 13, with occupations involving Administrative(clerical),protective(Services), Sales, Professional Speciality, Executive Managerical and Tech Support, and aged less than 34 years have a **95.16%** chance of earning more than $50k per year.
2. People who have a marital status as Married, have an education level less than 13, with occupations involving Administrative(clerical),protective(Services), Sales, Professional Speciality, Executive Managerical and Tech Support, and working for more than 8 hours per week in the Federal or Local Government or Private sector have a **95.39%** chance of earning more than 50k per year.
3. People with a marital status as Married, have an education level less than 13, occupation in Private House Services, Handlers(Cleaners), Farming(fishing), and aged more than 37 years old, have a **91.98%** chance of having an income greater than 50k per year.

6. CONCLUSION

As per the analysis conducted, it can be concluded that the most important predictor variables involved in this case are:

1. Marital Status
2. Education Level
3. Occupation
4. Age

The tree pruned to 5 splits is shown in the screenshot below:



Using these parameters in the same order of preference, we can target a customer base using the marketing campaign that has an annual income of more than $50k and hence is more likely to purchase the products from PineApple Inc.