PREDICTIVE MODELING PROJECT PRESENTATION

ANAND TAMHANKAR | SHAN XU | SHRUTHI TADAMIRI | URVI KOHLI

AGENDA

- Business Problem
- SEMMA
- Business Value
- Recommendations/Conclusions

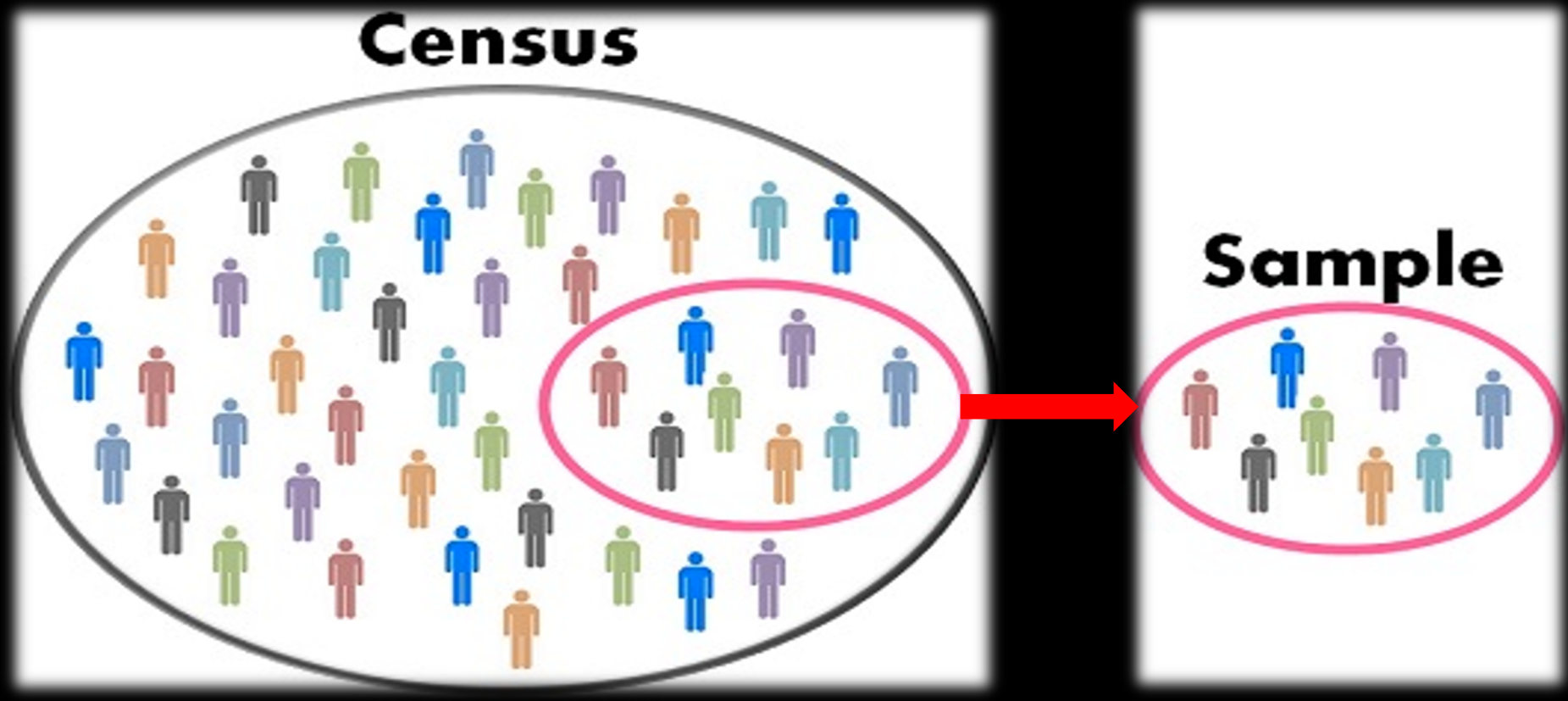
BUSINESS PROBLEM

Scenario: We are a team two Digital Marketing Managers and two Data Analysts at Apple Inc.

Problem to be solved:

- Limited budget for advertising on social media channels
- Targeted segment: customers with annual income > 50K
- Findings to be reported to VP – Marketing (Jose Cruz), on what variables are the most important in reaching the targeted segment

Sample

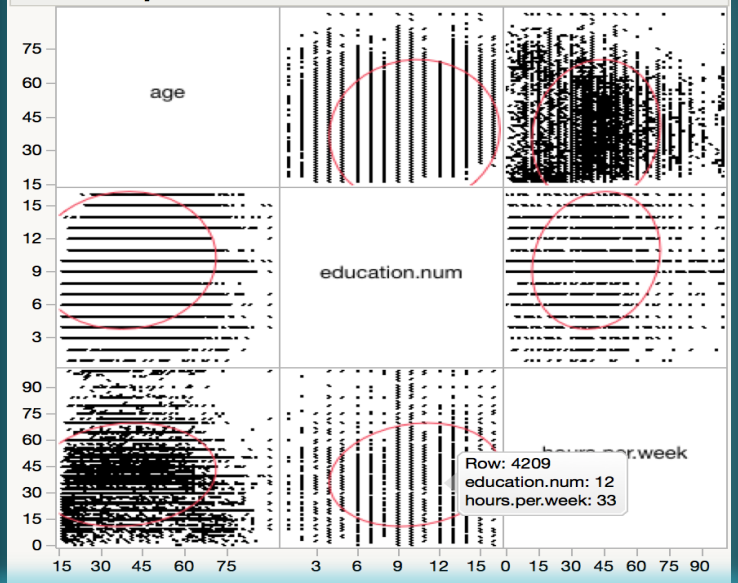


Explore

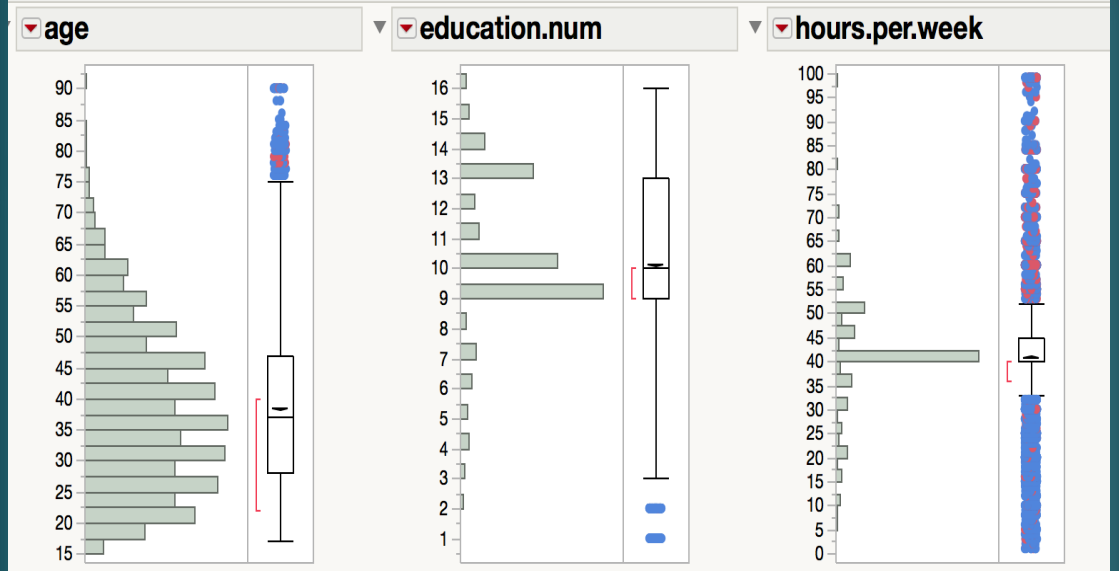
Correlations

	age	education.num	hours.per.week
age	1.0000	0.0435	0.1016
education.num	0.0435	1.0000	0.1525
hours.per.week	0.1016	0.1525	1.0000

Scatterplot Matrix



Distributions



Quantiles

100.0%	maximum	90
99.5%		76
97.5%		66
90.0%		57
75.0%	quartile	47
50.0%	median	37

Quantiles

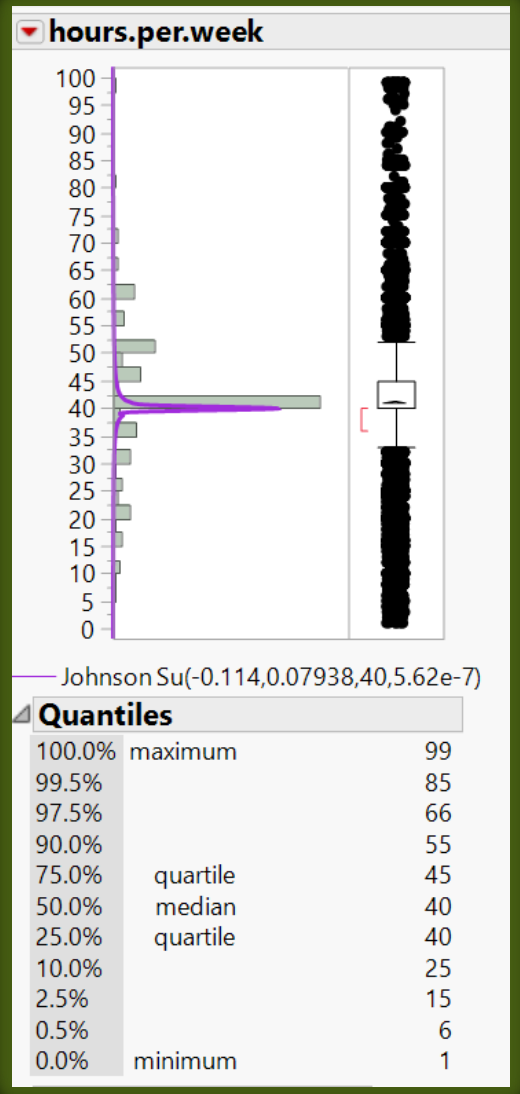
100.0%	maximum	16
99.5%		16
97.5%		15
90.0%		13
75.0%	quartile	13
50.0%	median	10

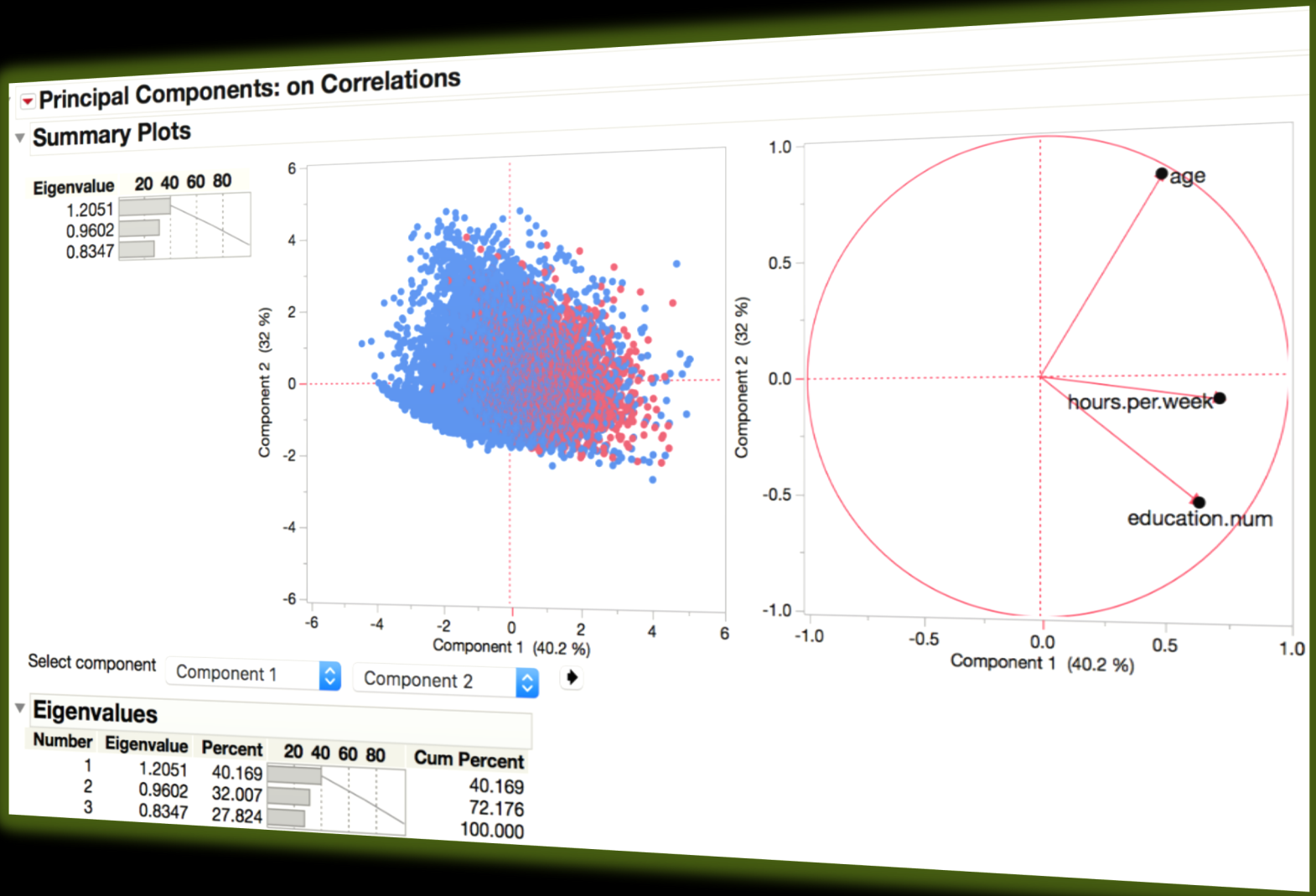
Quantiles

100.0%	maximum	99
99.5%		85
97.5%		66
90.0%		55
75.0%	quartile	45
50.0%	median	40

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

Modify


hours.per.week
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
Johnson Su(-0.114,0.07938,40,5.62e-7)
Quantiles
100.0% maximum 99
99.5% 85
97.5% 66
90.0% 55
75.0% quartile 45
50.0% median 40
25.0% quartile 40
10.0% 25
2.5% 15
0.5% 6
0.0% minimum 1


Principal Components: on Correlations
Summary Plots
Eigenvalue 20 40 60 80
1.2051
0.9602
0.8347
Component 2 (32 %)
Component 1 (40.2 %)
Select component
Component 1
Component 2
Eigenvalues
Number Eigenvalue Percent 20 40 60 80 Cum Percent
1 1.2051 40.169 40.169
2 0.9602 32.007 72.176
3 0.8347 27.824 100.000
age
hours.per.week
education.num
Component 2 (32 %)
Component 1 (40.2 %)

Model

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

Assess

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
Boosted Tree		0.3525	0.4844	0.3634	0.3417	0.2455	0.1731	18097
K Nearest Neighbors		.	.	.	.	.	0.1477	18097
Naive Bayes		.	.	.	.	.	0.2069	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
Boosted Tree		0.3680	0.5017	0.3546	0.3365	0.2423	0.1687	12065
K Nearest Neighbors		.	.	.	.	.	0.1728	12065
Naive Bayes		.	.	.	.	.	0.2027	12065

Small variation in values between Training and Validation sets = no over fitting!

Low misclassification rate and higher R Square

BUSINESS VALUE ANALYSIS

Base Scenario: Marketing Campaign (Random Selection)

Total Budget - **$100,000**

Cost per person - **$1**

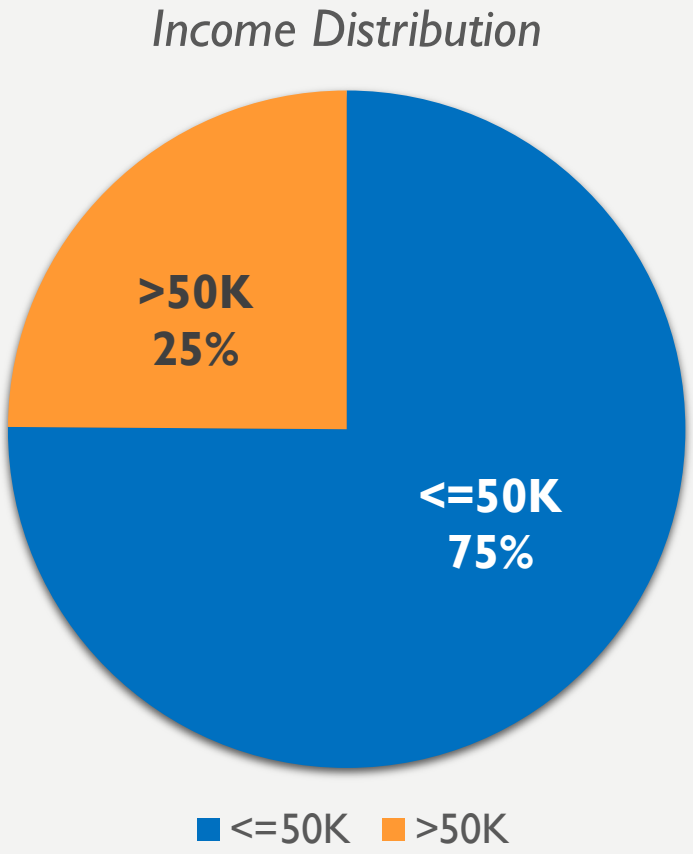
Total reach – **100,000 people**

% of selecting person with income $ >50K = 7508 / 30162 = **24.9%**

Revenue per customer - **$1000**

Conversion rate – **1%**

Expected Revenue - **$ 249,000**



BUSINESS VALUE ANALYSIS

Now, lets say we develop a model with a confusion Matrix like this

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

BUSINESS VALUE ANALYSIS

With a new model

Total reach – **100,000 people**

% of selecting person with income $ >50K = 1742 / 30162 = **83.2%**

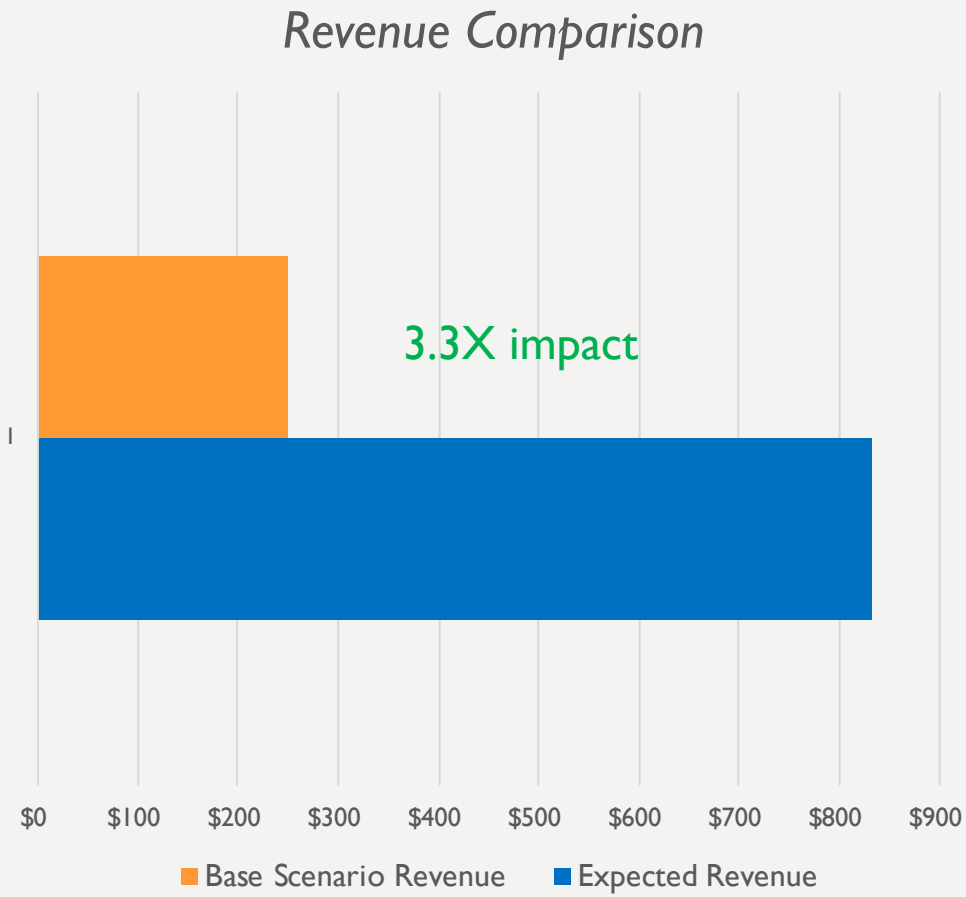
Revenue per customer - **$1000**

Conversion rate – **1%**

Expected Revenue - **$832,000**

Base Scenario Revenue - **$249,000**

Impact – 3.3x



RECOMMENDATIONS

Decision Tree Model

Column Contributions

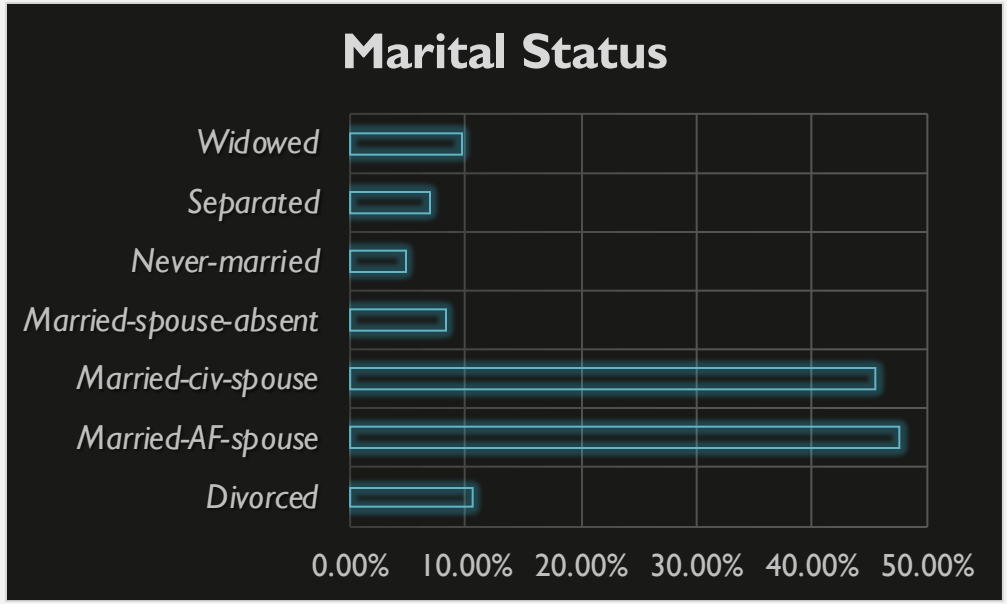
Term	Number of Splits	G^2	Portion
marital.status	1	3842.45803	0.5044
education.num	10	1853.43377	0.2433
age	17	846.498361	0.1111
occupation	10	591.380934	0.0776
hours.per.week	7	272.615768	0.0358
sex	6	77.8015884	0.0102
workclass	5	76.0769102	0.0100
relationship	3	43.0572348	0.0057
race	1	14.4686434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

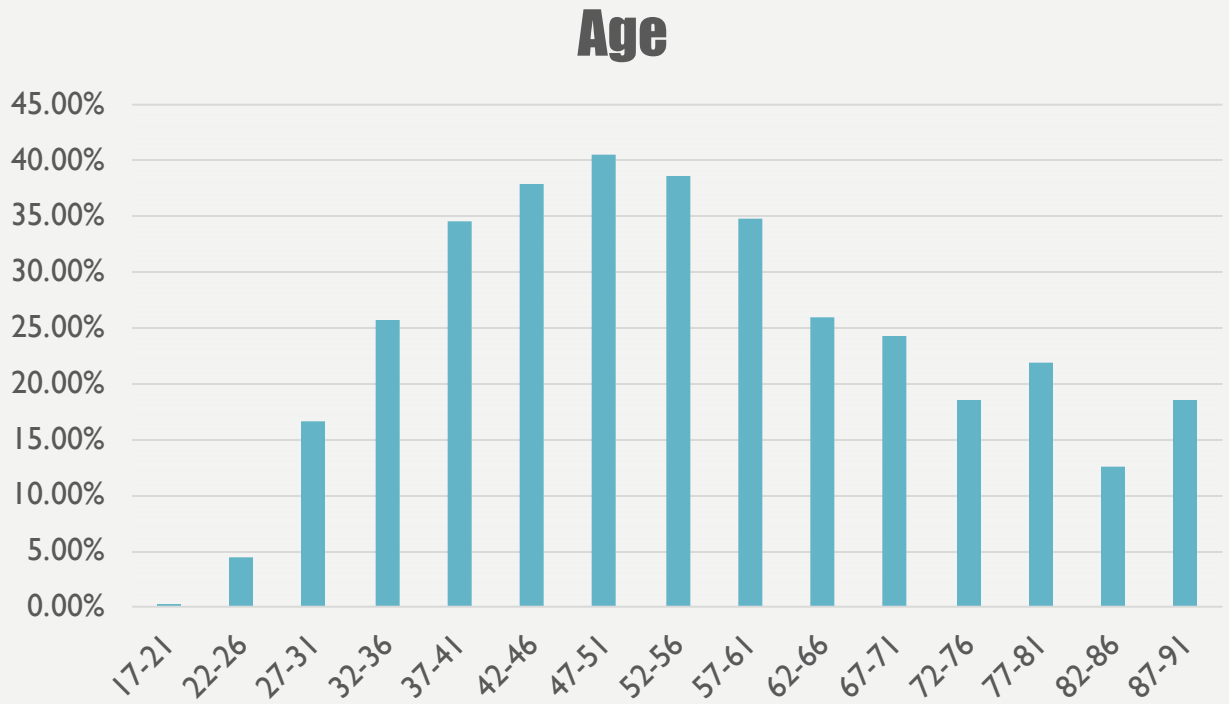
Effect Summary

Source	LogWorth	PValue
education.num	142.891	0.00000
occupation	61.985	0.00000
age	47.492	0.00000
relationship	41.899	0.00000
hours.per.week	37.705	0.00000
sex	23.292	0.00000
workclass	11.641	0.00000
marital.status	10.713	0.00000
native.country	0.856	0.13939
race	0.655	0.22112

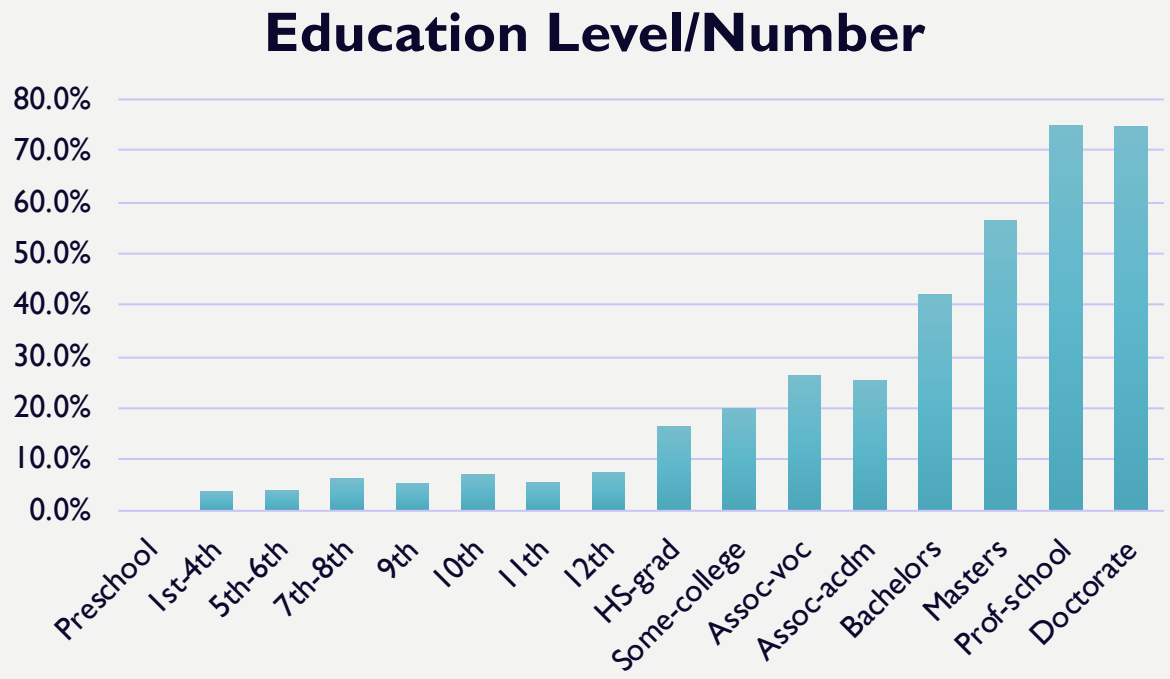
Logistic Regression Model

RECOMMENDATIONS

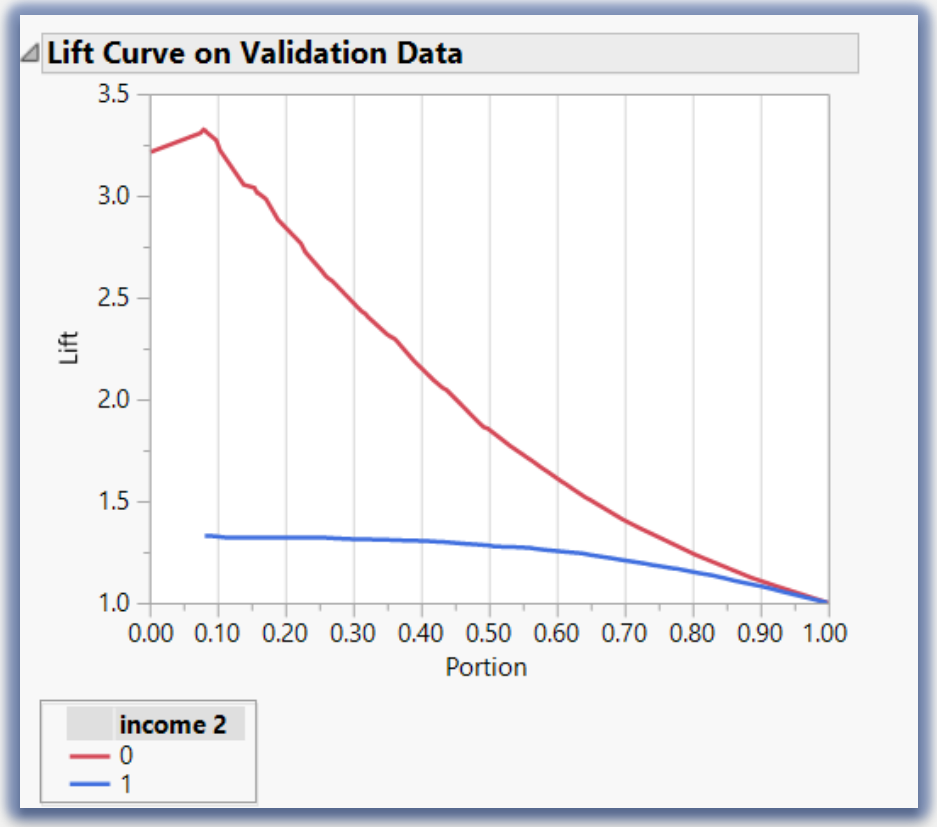




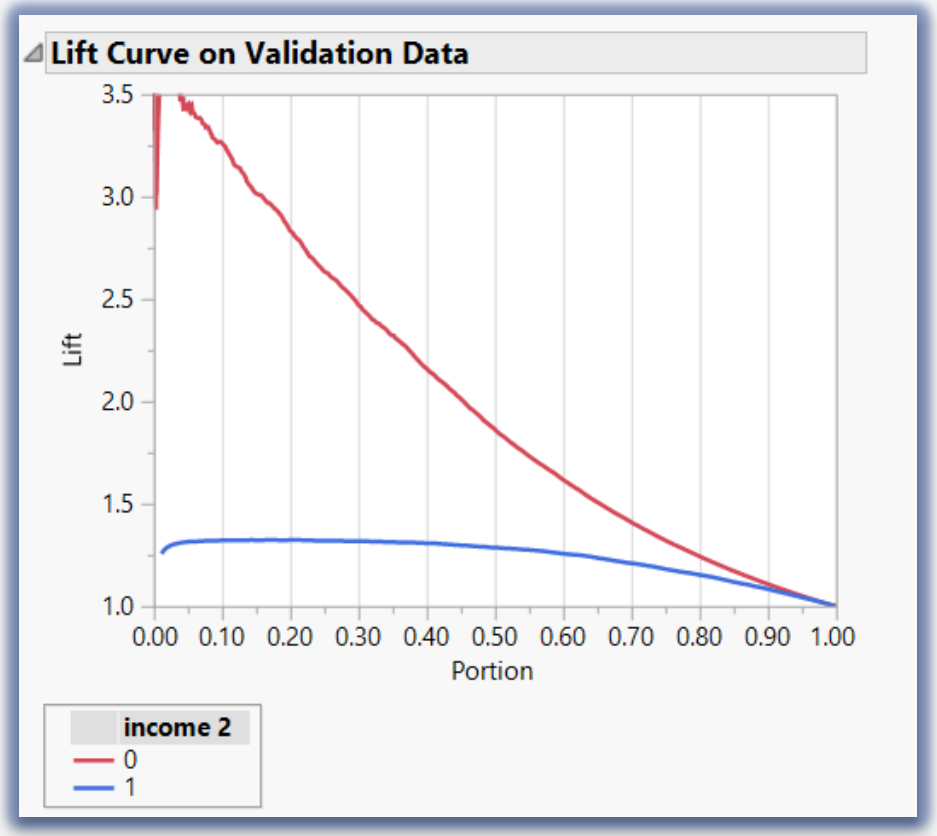
RECOMMENDATIONS



RECOMMENDATIONS



Decision Tree Model



Logistic Regression Model

RECOMMENDATIONS

Based on the models ran by our analysts, we can say that these four variables are important in targeting the desired customer segment through the marketing campaign

- Marital Status
- Education Number
- Occupation
- Age

Thank You