Identifying Prospective Customers For Visa Applications Agency

A Predictive Analysis on H-1B Data

Stats Squad

Anuj Vengurlekar, Mingwei Zhang, Kaitlin McGahie, Yingzi Ma

Agenda

- Introduction
- Data Characteristics and Challenges Faced
- Initial Approach and final study questions
- Supervised Learning
  - Logistic Regression
  - Decision Tree
- Unsupervised Learning
  - Cluster Analysis
- Conclusion
- References and Data Sources

Introduction

- US Citizenship and Immigration Services
- H-1B Visa Application & Process

CASE 1:

John, a full-time Senior Analyst at Deloitte Consulting. His annual salary is $102,461 and he filed the H1B application with a visa application agency. Did he get certified?

CASE 2:

Emily, a full-time Executive at Cisco System. Her annual salary is $239,159 and she filed the H1B application with a visa application agency. Did she get certified?

Unfortunately, they both got denied!

Data Characteristics

Data Source: US Office of Labour Certification (OFLC)

Data-set: 600,000 applicants

40 variables reduced down to 14 variables

Variables include: Employment Details, Details of Employee, Details of Company, Dependent?, Type of Application

Out of 14 variables: 3 quanititative and 12 qualitative

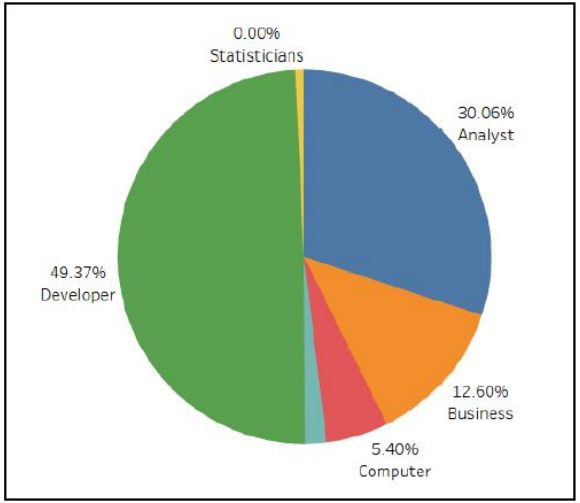
Challenges Faced

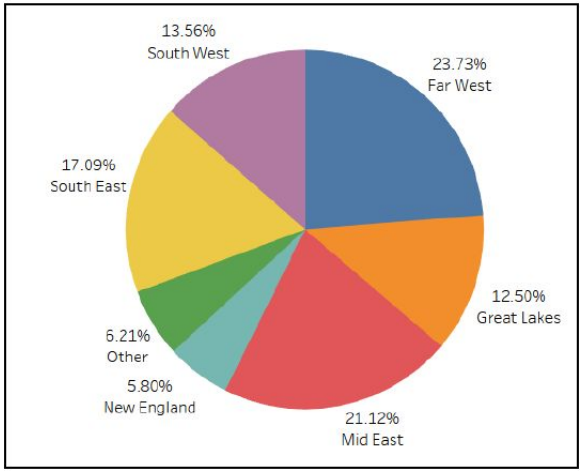
- Data Cleaning

- Eliminating Incomplete Applications / Missing Data:
  - Calculating Variables
    - Number of applicants in a company
    - Wage
- Refining Data Using Different Sources:
  - Regions
  - Sector
  - Job Function and Domain

Visual Summary of Applicants

Sector	# of Records	Percentage
FinInsur	12024	5.05%
Information	13994	5.88%
Manufacturing	8675	3.65%
Other	10416	4.38%
ProfScienTech	187356	78.74%
Retail Trade	5478	2.30%
Grand Total	237943	

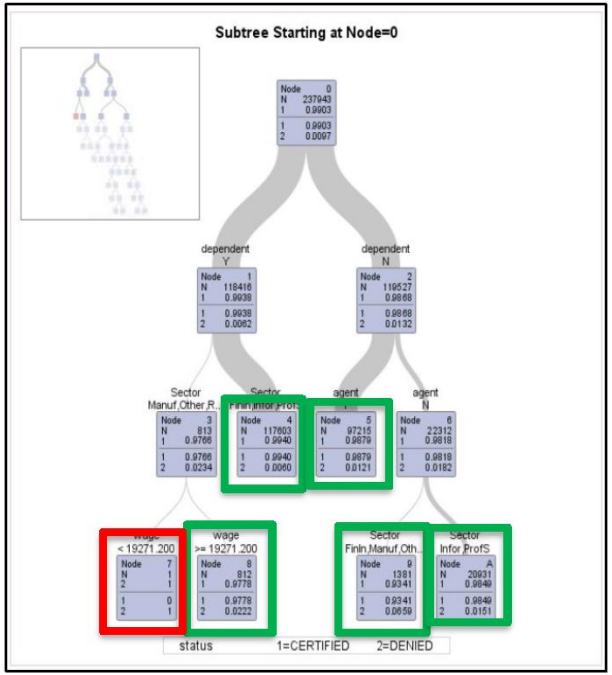




Initial Approach and Issues Found

- Can we predict whether or not someone's work authorization request will be certified?
- Not a useful question to ask:

Status	No of Apps	Percentage
Certified	235,635	99%
Denied	2,309	1%
Total Applications	**237,944**	



Final Study Questions

- What predictors influence the likelihood that a person/company uses an agent during the work authorization application?
- Do factors like industry, job function, or domain have an effect on whether or not an agency is used for a work authorization application?
- Can we predict which demographic group is most likely to use an agent?

Logistic Regression

- Expression of the logistic regression model
- 11 interaction terms (“*wage - Sector*”, “*wage - Region*”)
- 16 out of 23 statistically significant variables

$$P(Y = \text{“Hire an agent”} \mid X) = p = \frac{\exp(E)}{1+\exp(E)}$$

Where

$$\begin{aligned} E = {} & \beta_0 + \beta_{wage}X_1 + \beta_{Sector\_FinIn}X_2 + \beta_{Sector\_Info}X_3 \\ & + \beta_{Sector\_Manuf}X_4 + \beta_{Sector\_Other}X_5 \\ & + \beta_{Sector\_ProfS}X_6 + \beta_{Region\_FW}X_7 \\ & + \beta_{Region\_GL}X_8 + \beta_{Region\_ME}X_9 \\ & + \beta_{Region\_NE}X_{10} + \beta_{Region\_Ot}X_{11} \\ & + \beta_{Region\_SE}X_{12} + \beta_{wage}\beta_{Region\_FW}X_{13} \\ & + \beta_{wage}\beta_{Region\_GL}X_{14} + \beta_{wage}\beta_{Region\_ME}X_{15} \\ & + \beta_{wage}\beta_{Region\_NE}X_{16} + \beta_{wage}\beta_{Region\_Ot}X_{17} \\ & + \beta_{wage}\beta_{Region\_SE}X_{18} + \beta_{wage}\beta_{Sector\_FinIn}X_{19} \\ & + \beta_{wage}\beta_{Sector\_Info}X_{20} + \beta_{wage}\beta_{Sector\_Manuf}X_{21} \\ & + \beta_{wage}\beta_{Sector\_Other}X_{22} + \beta_{wage}\beta_{Sector\_Prof}X_{23} \end{aligned}$$

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	4.4627	0.4467	99.8155	<.0001
wage		1	-4.2E-6	4.741E-6	0.7862	0.3752
Sector	FinIn	1	-0.5069	0.5082	0.9946	0.3186
Sector	Infor	1	-2.0777	0.4507	21.2551	<.0001
Sector	Manuf	1	-1.5654	0.5028	9.6943	0.0018
Sector	Other	1	-3.7225	0.4555	66.7788	<.0001
Sector	ProfS	1	-5.4511	0.4439	150.8082	<.0001
Sector	Retai	0	0	.	.	.
Region	FW	1	-0.6652	0.0667	99.4794	<.0001
Region	GL	1	0.2467	0.0804	9.4202	0.0021
Region	ME	1	-0.2528	0.0687	13.5426	0.0002
Region	NE	1	0.1505	0.1084	1.9251	0.1653
Region	Ot	1	-0.3408	0.1057	10.4031	0.0013
Region	SE	1	-0.6015	0.0724	69.0546	<.0001
Region	SW	0	0	.	.	.
wage*Sector	FinIn	1	2.635E-7	5.425E-6	0.0024	0.9613
wage*Sector	Infor	1	-9.02E-6	4.746E-6	3.6163	0.0572
wage*Sector	Manuf	1	0.000010	5.439E-6	3.5594	0.0592
wage*Sector	Other	1	0.000022	4.915E-6	20.8411	<.0001
wage*Sector	ProfS	1	0.000018	4.69E-6	14.2727	0.0002
wage*Sector	Retai	0	0	.	.	.
wage*Region	FW	1	0.000012	8.753E-7	177.5008	<.0001
wage*Region	GL	1	-1.49E-6	1.151E-6	1.6851	0.1942
wage*Region	ME	1	6.199E-6	9.439E-7	43.1268	<.0001
wage*Region	NE	1	3.016E-6	1.439E-6	4.3915	0.0361
wage*Region	Ot	1	4.676E-6	1.476E-6	10.0349	0.0015
wage*Region	SE	1	9.627E-6	1.025E-6	88.2244	<.0001
wage*Region	SW	0	0	.	.	.

Logistic Regression

Hire an agent?

Helpful in making predictions:

Sector_Information
Sector_Manufacturing
Sector_Other
Sector_Professional Scientific Technology
Region_Farwest
Region_Great Lakes
Region_Mideast
Region_Other
Region_SouthEast
Wage*Sector_others
Wage*Sector_Professional Scientific Technology
Wage*Region_Farwest
Wage*Region_Mideast
Wage*Region_New England
Wage*Region_Other
Wage*Region_Southeast

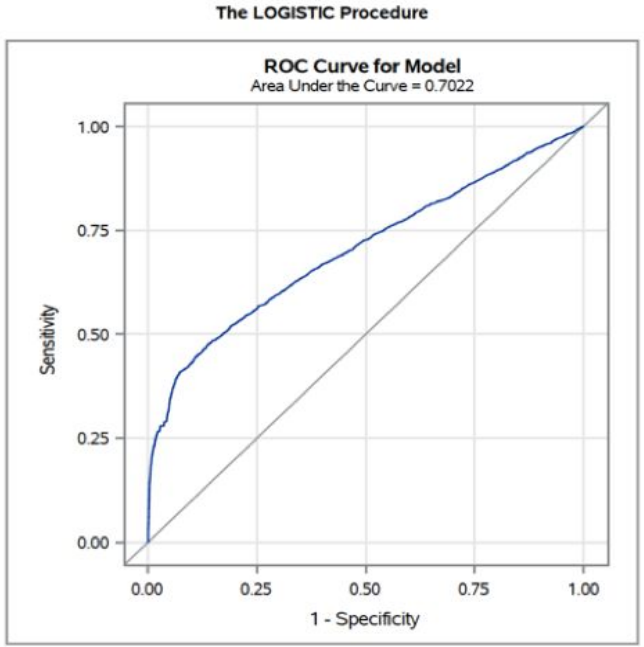
Not helpful in making predictions:

Wage
Sector_Finance, Insurance
Region_New England
Wage*Sector_Finance, Insurance
Wage*Sector_Information
Wage*Sector_Manufacturing
Wage*Region_Great Lakes

Logistic Regression

Check Model Utility

- AUC = 0.7022
- 64% correctly classified at 0.52 or 0.54 cutoff



Classification Table

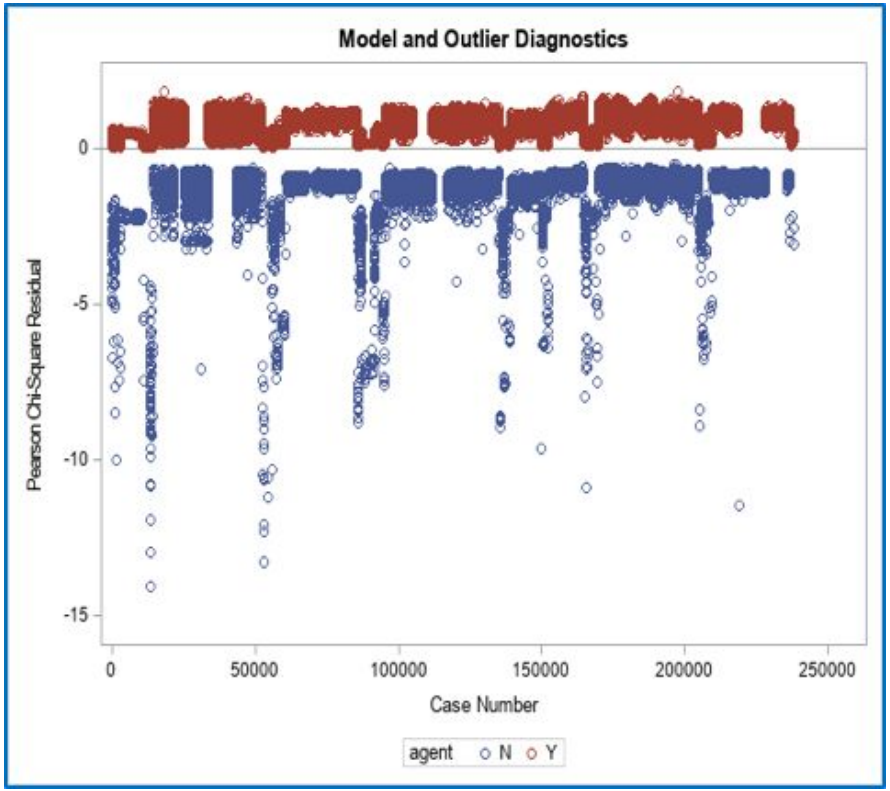
Prob Level	Correct		Incorrect		Percentages				
	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.220	147E3	0	90895	0	61.8	100.0	0.0	38.2	.
0.240	147E3	2	90893	2	61.8	100.0	0.0	38.2	50.0
0.260	147E3	2	90893	2	61.8	100.0	0.0	38.2	50.0
0.280	147E3	5	90890	12	61.8	100.0	0.0	38.2	70.6
0.300	147E3	31	90864	47	61.8	100.0	0.0	38.2	60.3
0.320	147E3	158	90737	124	61.8	99.9	0.2	38.2	44.0
0.340	147E3	241	90654	270	61.8	99.8	0.3	38.2	52.8
0.360	147E3	504	90391	484	61.8	99.7	0.6	38.1	49.0
0.380	145E3	1972	88923	2024	61.8	98.6	2.2	38.0	50.7
0.400	144E3	3479	87416	2909	62.0	98.0	3.8	37.8	45.5
0.420	142E3	6627	84268	5527	62.3	96.2	7.3	37.3	45.5
0.440	138E3	10678	80217	9123	62.5	93.8	11.7	36.8	46.1
0.460	131E3	18017	72878	15558	62.8	89.4	19.8	35.7	46.3
0.480	121E3	30098	60797	26446	63.3	82.0	33.1	33.5	46.8
0.500	111E3	41152	49743	36433	63.8	75.2	45.3	31.0	47.0
0.520	101E3	51152	39743	46026	64.0	68.7	56.3	28.2	47.4
0.540	93109	59227	31668	53939	64.0	63.3	65.2	25.4	47.7
0.560	84944	65883	25012	62104	63.4	57.8	72.3	22.7	48.5
0.580	78151	71959	18936	68897	63.1	53.1	79.2	19.5	48.9

Logistic Regression

Check Model Validity

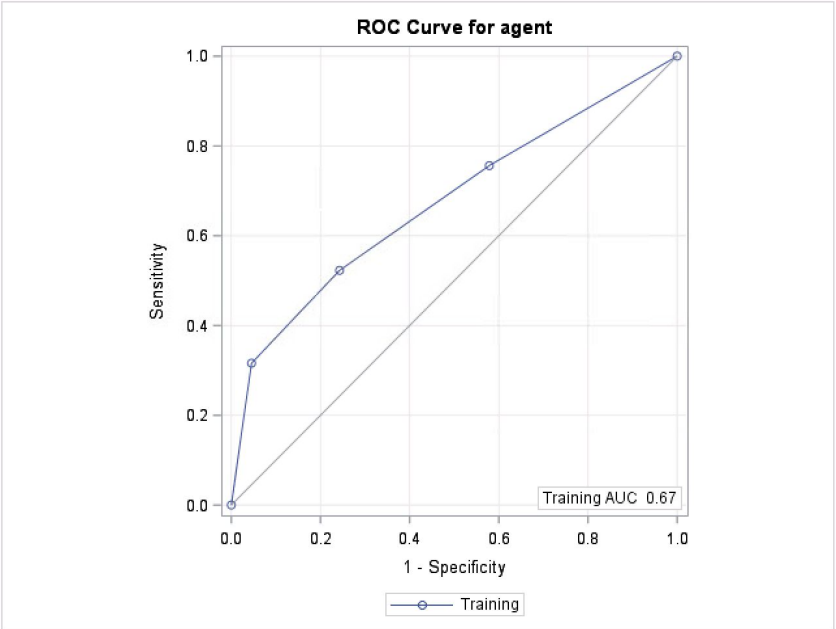
- Residual plot has pattern
- Reject null hyphthesis in HL test
- Will not further explain the model nor use to make predictions
- Conclusions and Limitations

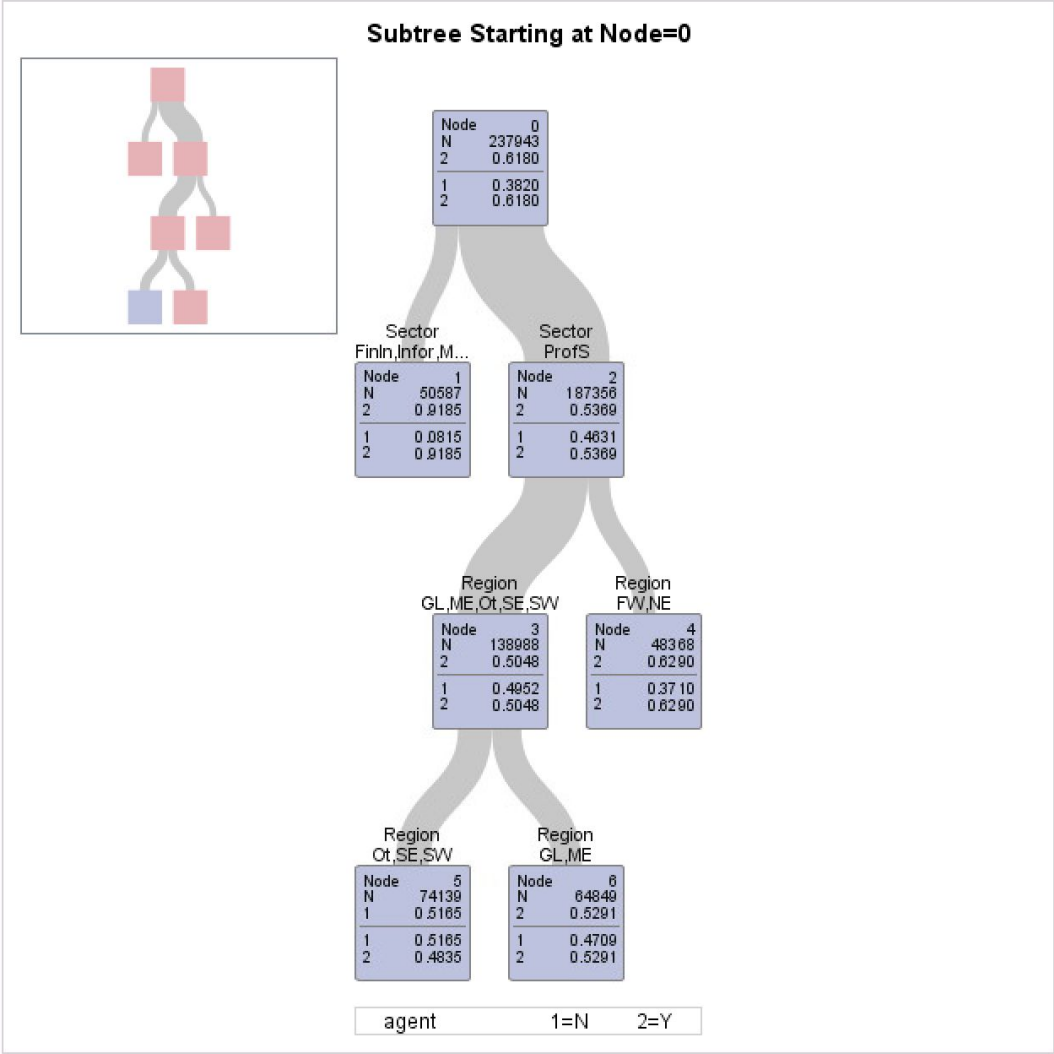
Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
1059.8997	8	<.0001



Decision Tree 1

Tree 1 - Predictors: region, sector



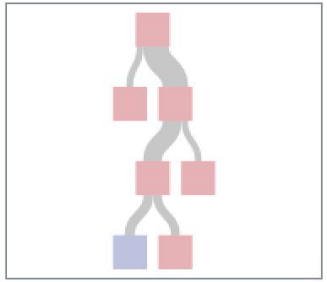


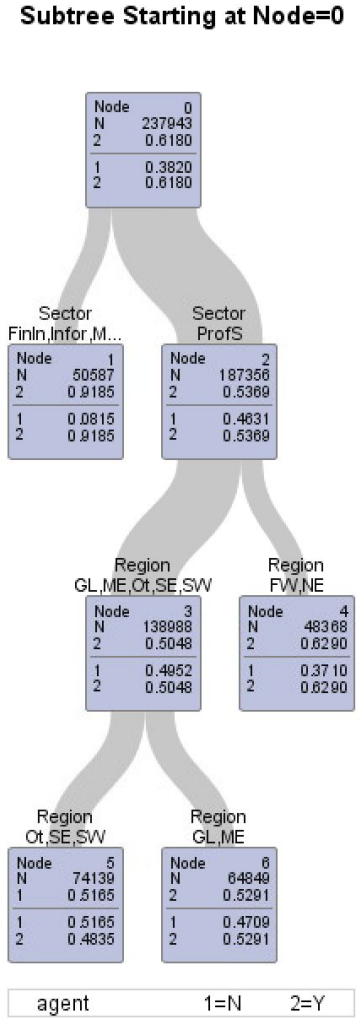
Decision Tree 1

Tree 1 - Predictors: region, sector

Model-Based Confusion Matrix			
Actual	Predicted N	Predicted Y	Error Rate
N	38293	52602	0.5787
Y	35846	111202	0.2438

Model-Based Fit Statistics for Selected Tree								
N Leaves	ASE	Mis-class	Sensitivity	Specificity	Entropy	Gini	RSS	AUC
4	0.2091	0.3717	0.7562	0.4213	0.8632	0.4181	99490.2	0.6748



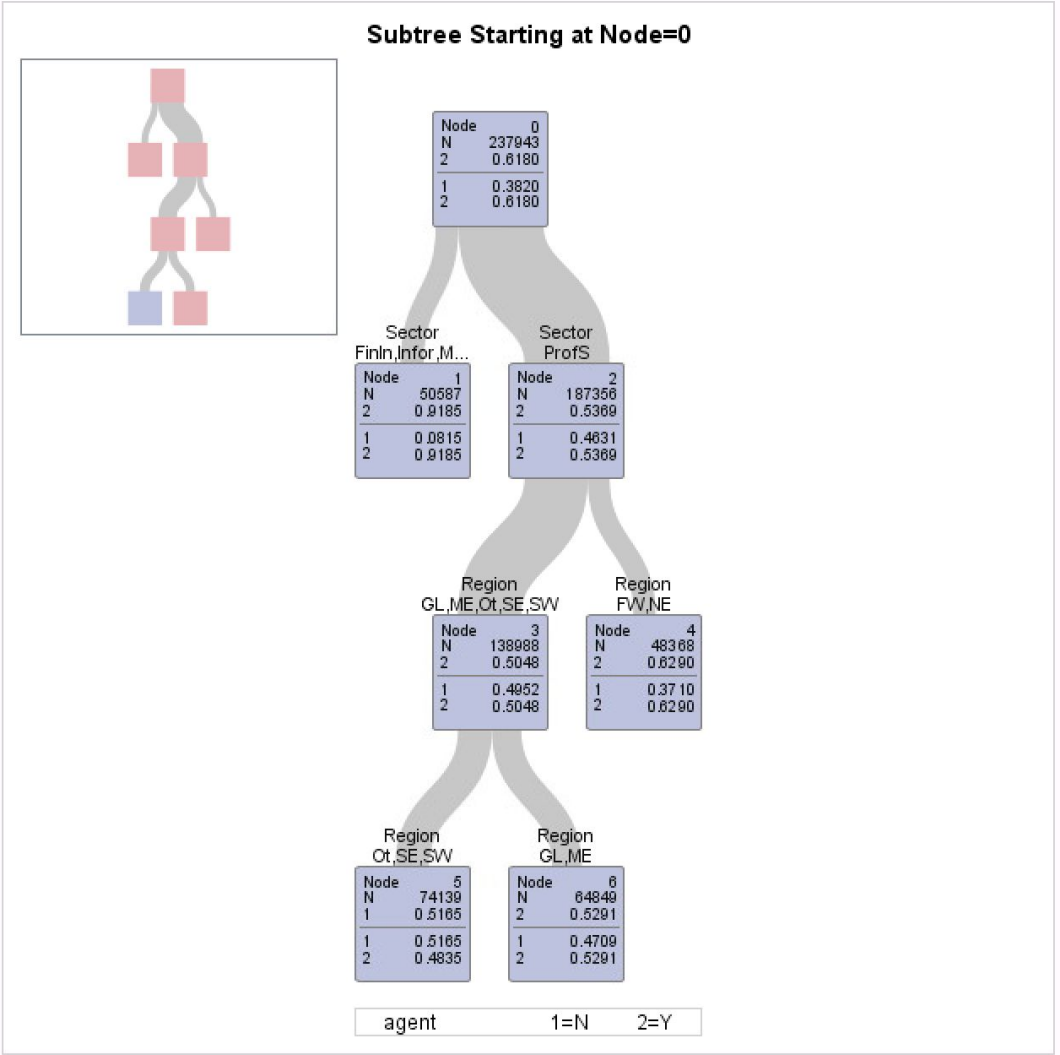


Decision Tree 1

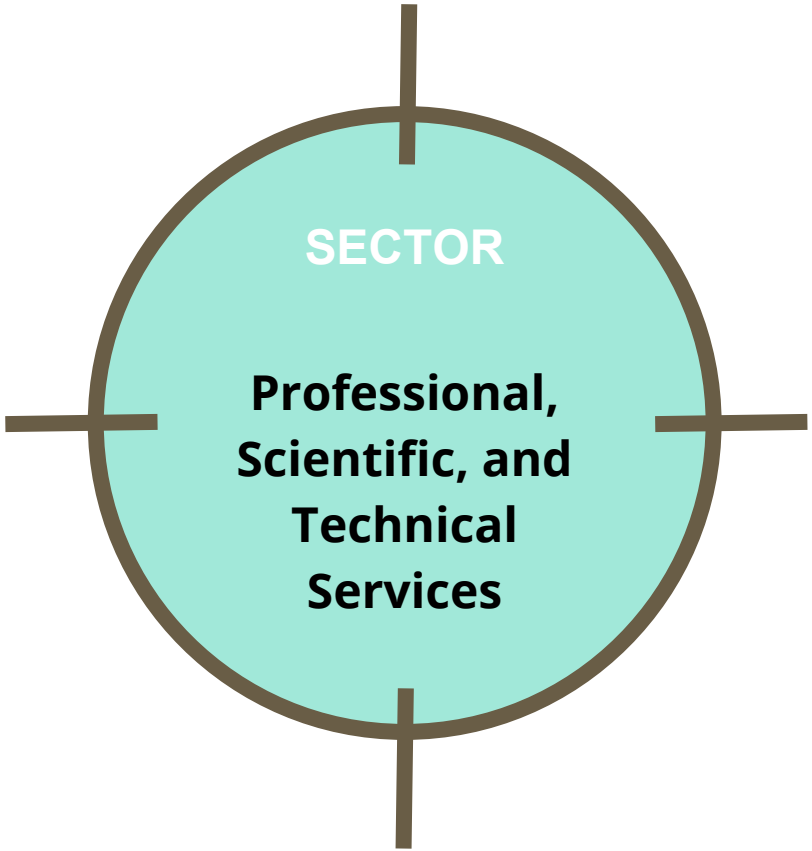
Tree 1 - Predictors: region, sector

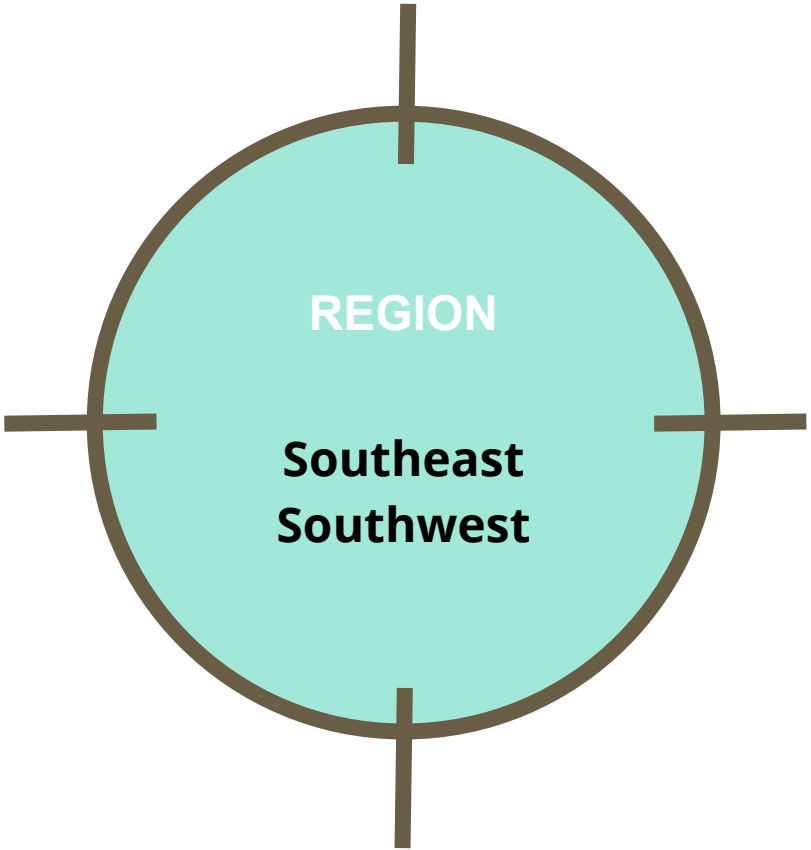
Findings:

- Applicants from sectors other than Professional, Scientific, and Technical Services → agent ('Y') (92%)
- Applicants from Professional, Scientific, and Technical Services
  - Far West and New England → agent ('Y')
  - Southeast, Southwest & other → agent ('N')



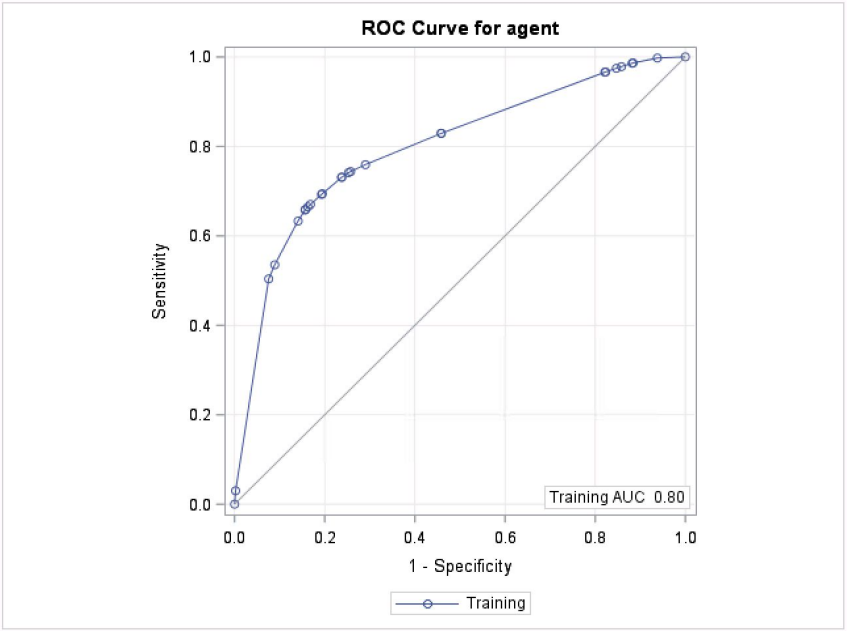
Business Insights from Decision Tree 1

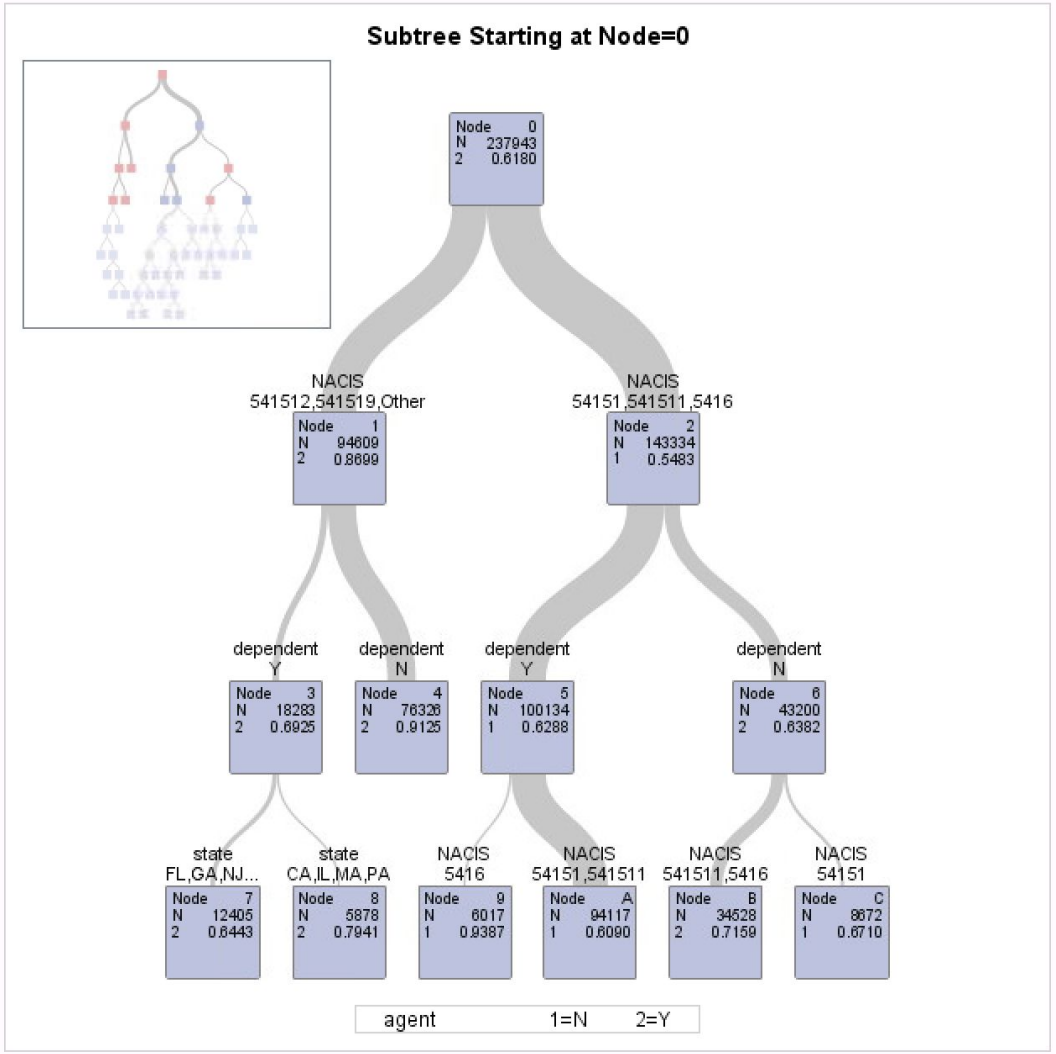




Decision Tree 2

Tree 2 - Predictors: NAICS, dependent, state



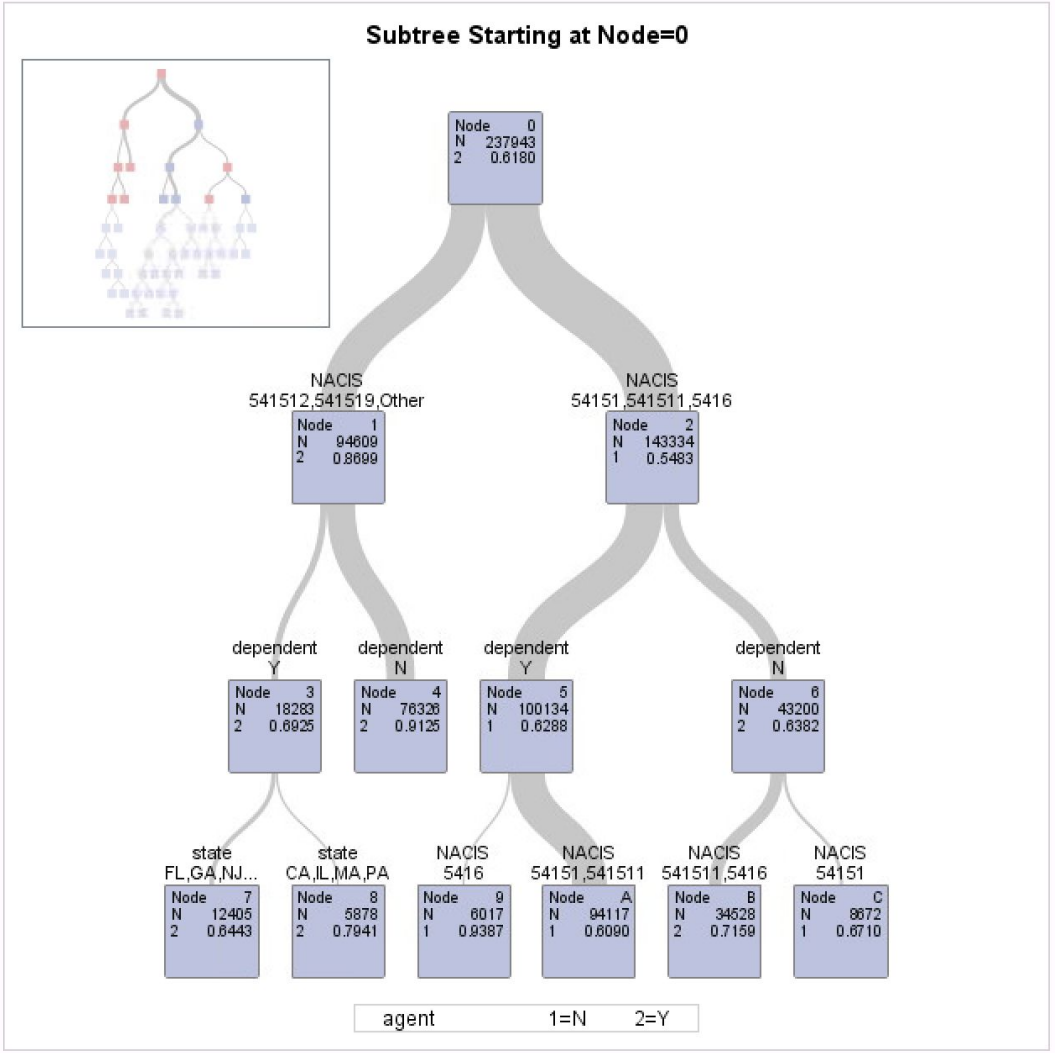


Decision Tree 2

Tree 2 - Predictors: NAICS, dependent, state

Model-Based Confusion Matrix			
Actual	Predicted N	Predicted Y	Error Rate
N	67947	22949	0.2525
Y	38055	108992	0.2588

Model-Based Fit Statistics for Selected Tree								
N Leaves	ASE	Mis-class	Sensitivity	Specificity	Entropy	Gini	RSS	AUC
25	0.1705	0.2564	0.7412	0.7475	0.7374	0.3410	81130.6	0.7987

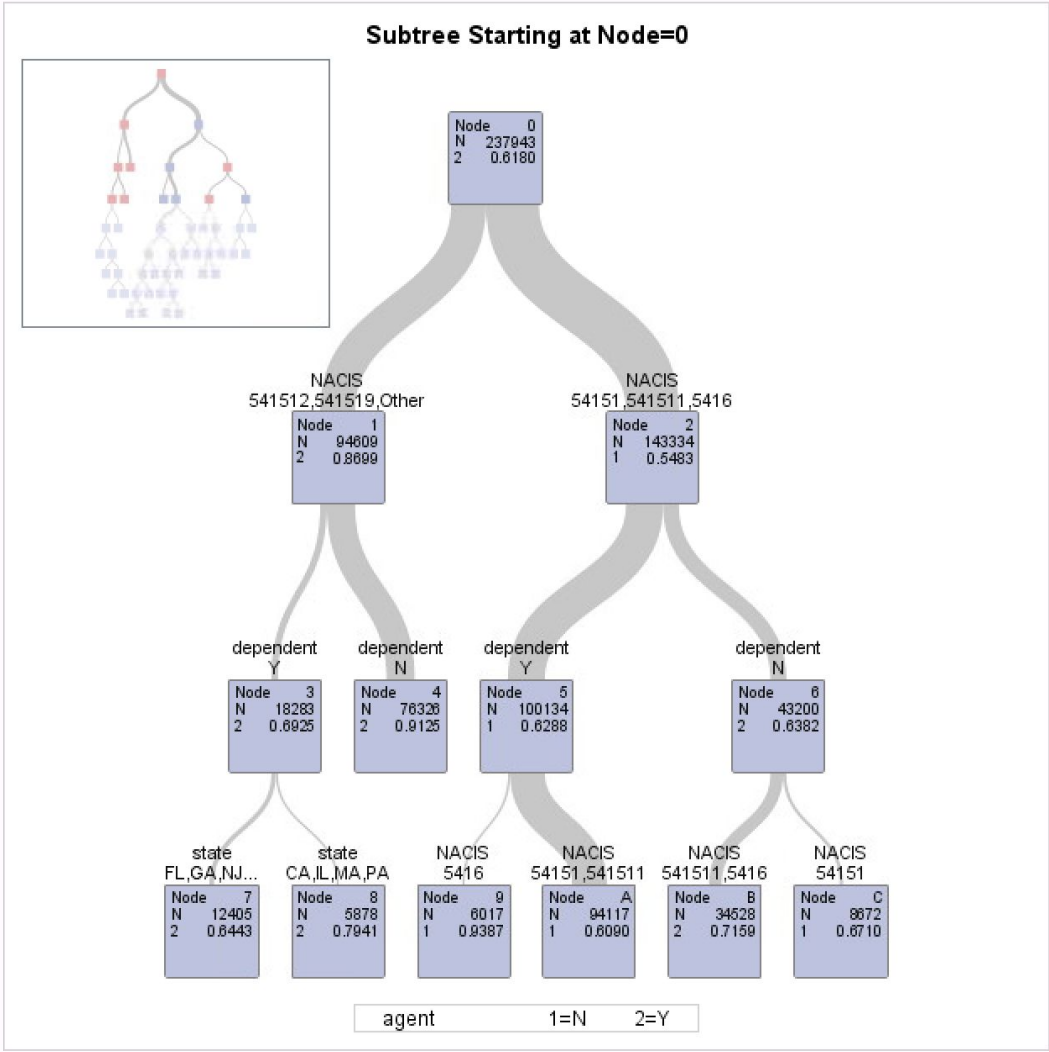


Decision Tree 2

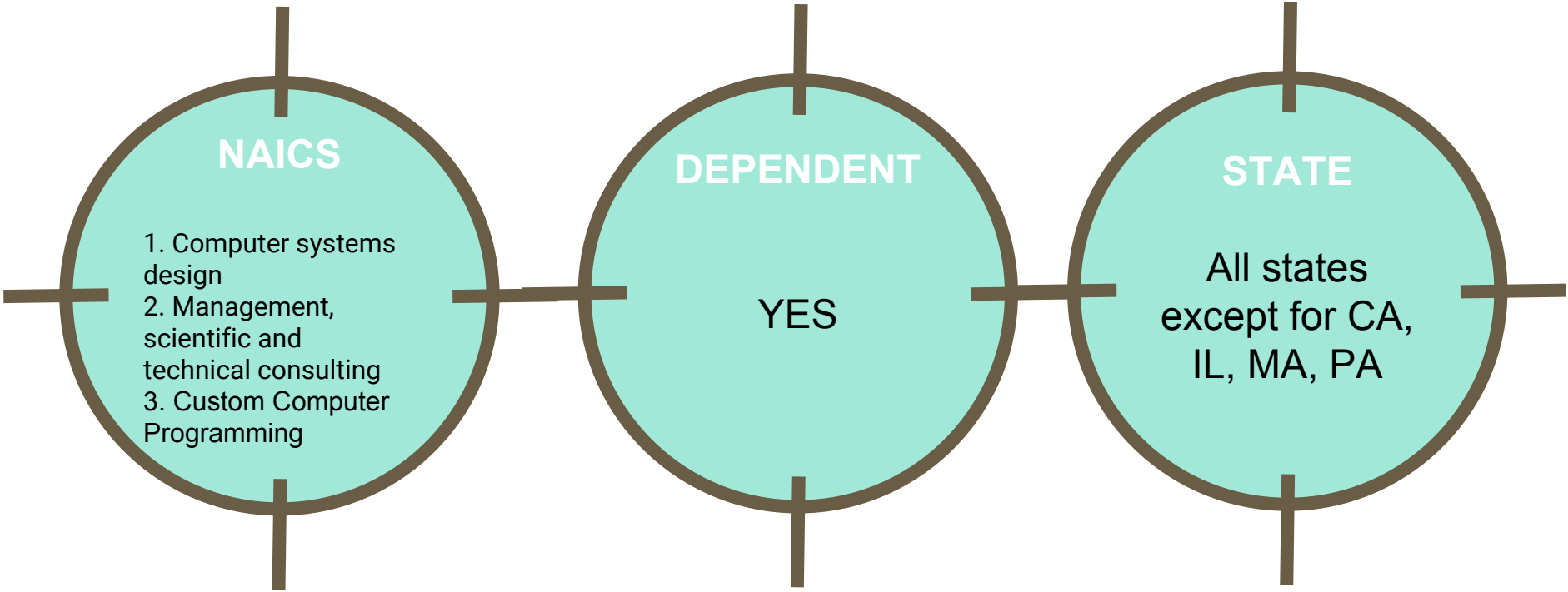
Tree 2 - Predictors: NAICS, dependent, state

Findings:

- NAICS (541512, 541519 & other) → agent ('Y') (87%)
  - dependent ('Y') → CA, IL, MA, PA → agent ('Y')
- NAICS (54151, 541511, 5416) → agent ('N') (55%)
  - dependent ('Y') → agent ('N')
  - dependent ('N') → agent ('Y')



Business Insights from Decision Tree 2



Limitations

- No statistics basis/assumptions for decision trees
- Prone to overfitting - highly dependent on the dataset
- Random forest is a more robust choice

Cluster Analysis

3 Quantitiatve variables considered for clustering

4 Clusters

Complete Linkage method

Did not have any quantitative variables left so we compared with qualitative variables

Cluster Results

CLUSTER	N Obs	Variable	Label	N	Mean
1	467	duration	duration	467	1069.51
		wage	wage	467	89603.58
		Companyapps	Companyapps	467	2099.87
2	57	duration	duration	57	1094.86
		wage	wage	57	135780.93
		Companyapps	Companyapps	57	855.7894737
3	550	duration	duration	550	1062.90
		wage	wage	550	60315.98
		Companyapps	Companyapps	550	1748.33
4	4	duration	duration	4	1094.75
		wage	wage	4	195837.00
		Companyapps	Companyapps	4	2156.50

Cluster Vs Regions

Cluster	Region	
1	Far West	29.34%
	Great Lakes	9.85%
	Mid East	22.27%
	New England	7.71%
	Other	6.42%
	South East	13.28%
	South West	11.13%
2	Far West	77.19%
	Mid East	17.54%
	New England	1.75%
	South West	3.51%
3	Far West	11.09%
	Great Lakes	17.09%
	Mid East	21.09%
	New England	5.45%
	Other	7.45%
	South East	22.18%
	South West	15.64%
4	Far West	75.00%
	Mid East	25.00%

Cluster Vs Roles

Cluster	Role	
1	Analyst	31.26%
	Business	11.13%
	Computer	4.07%
	Database	1.50%
	Developer	51.61%
	QA	0.43%
2	Analyst	10.53%
	Business	3.51%
	Computer	8.77%
	Database	1.75%
	Developer	75.44%
3	Analyst	34.00%
	Business	13.09%
	Computer	3.64%
	Database	2.55%
	Developer	46.18%
	QA	0.55%
4	Computer	100.00%

Clusters

High Skilled Techies

Key Traits:

- Experienced Professionals in developer roles
- Located in California and Washington region
- Mean salary of 140k
- large organizations

Corporate Specialists

Key Traits:

- Experienced Professionals mix of roles
- Big cities: SF, New York, Dallas, Chicago
- Mean salary of 85k
- Large to Medium size organizations

High-level Bosses

Key Traits:

- VP/Director Level Employees
- Mean salary of $220k +
- Computer Architecture role

Corporate Associates

Key Traits:

- Low median salary $55k to $60k
- Applicants new in their career
- Big cities: SF, NEw York, Dallas, Chicago
- Large to Medium size organizations

Limitations

- Clustering is subjective and requires domain knowledge
- We referred to mean values from the clusters and homogeneity within the clusters has not been thoroughly checked before drawing conclusions
- Clustering is based on smaller sample size due to computational constraints
- Clustering is specific to the dataset and hence generalizing these results is difficult

Concluding Remarks

- Huge market opportunity for visa agencies (600,000 applicatants per year)
- Visa Agencies can explore targeting
- Professional, Scientific and Technical Sector and Applicants with H1B dependent
- Southwest/ Southeast regions and states other than CA, IL MA and PA
- 4 distinct Cluster groups to identified amongst the applicants

Future Scope

- Performance of the models can be tested on data from different Fincal years
- Perform clustering on multiple samples so see if we have consistent clustering
- Analysis on results of actual H1B certifications

References and Data Sources

Neil Liberman, 2017. Towards Data Science. “Decision Trees and Random Forests”. https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991

Kramer AA, Zimmerman JE, 2007. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. https://www.ncbi.nlm.nih.gov/pubmed/17568333

North America Industry Classification System https://www.census.gov/cgi-bin/sssd/naics/naicsrch

Bureau of Economic Analysis https://www.bea.gov/data/economic-accounts/regional

Step by Step H1B visa applications guide https://www.immi-usa.com/h1b-application-process-step-by-step-guide/

U.S. Department of Labor https://www.dol.gov/

Thank You!

Questions?