
# Identifying Prospective Customers For Visa Applications Agency

A Predictive Analysis on H-1B Data

Anuj Vengurlekar | Mingwei Zhang | Kaitlin McGahie | Yingzi Ma

December 14, 2018

## TABLE OF CONTENTS

# I. Abstract:

Prior to the submission of an H-1B United States Visa application, a foreign individual must first be authorized to work in the United States. Many of these individuals choose to use an agent to aide in their application process. This analysis uses a comprehensive dataset,that contains thousands of observations across a number of categories related to these work authorization applications, to build predictive models that attempt to understand the relationship between these captured variables and whether or not an applicant or the company chooses to use an agent.

# II. Introduction:

Every year, the United States Citizenship and Immigration Services (USCIS) department oversees the collection and vetting of hundreds of thousands of U.S. visa applications. There are a number of different kinds of visas—nonimmigrant visas, for short-term or temporary visits to the U.S., and immigrant visas, for those who wish to immigrate to the U.S. permanently.

Of the temporary visas, the most common for skilled laborers is the H-1B. The H-1B visa grants a foreign worker a temporary stay (three years with the opportunity for an additional three-year renewal) while they are employed in a specialty position, generally a position that requires a degree of tertiary education, at a U.S. employer. In 1990, Congress set an annual limit on the number of new H-1B visas that could be granted each fiscal year. This cap has been reached year after year since nearly the beginning of the 21st century, and since H-1B visas are currently the most viable and only truly feasible way that international students and skilled foreign workers can stay in the United States long term, the administrative process for applicants has high stakes.

Prior to submitting an official H-1B application, applicants seeking a visa must first be authorized to work in the United States. It was initially the objective of this study to explore and understand the relationships and interactions between a number of variables—state of residence, industry of desired employment, the inclusion of a dependent, agent status, and salary—and the likelihood of U.S. work authorization. However, that data turned out to be highly skewed, because almost everybody who applies for work authorization is granted it. The following step--H-1B certification, is where the numbers thin out. Thus, in initial iterations of the primary analyses, decision trees resulted in “event” (or work authorization) classification of all observations.

In the interest of building and providing a more applicable model, it became the objective of this study to understand the relationship between the aforementioned variables and the likelihood that a given work authorization applicant uses an agent. Furthermore, the dataset we used contained far more observations than felt relevant to this team. Thus, professions outside the realm of business study applications were excluded from the model. The primary question being addressed became: what factors make a U.S. work authorization applicant more likely to use an agent, and to what extent? The following sections will explore this question, and will contain detailed descriptions of the variables used, the process by which our models were built, and the subsequent in-depth analysis and related conclusions our models lend themselves to.

## III. Data Characteristics:

The Office of Foreign Labor Certification (OFLC) generates program data, including data about H-1B visas. The disclosure data is updated annually and is available online. The primary aim was to familiarize ourselves with the data and understand the intricacies and features of the selected dataset. The dataset included information on over 600,000 applicants who applied for H-1B LCA in the 2017 fiscal year.

The raw data that was made available was messy and not immediately suitable for analysis. A number of data transformations were performed, which made the data more accessible for quick exploration.

The final list of variables used for analysis after applying transformations in the dataset are given in the **Appendix** along with visual summaries of key variables.

## IV. Challenges faced while working with the data-set

### Data Cleaning:

**Eliminating Incomplete Applications/Missing Data:**

We had to perform Data Cleaning by eliminating the data that had few missing values associated with it. This was done to ensure that the dataset taken into account was of good quality and did not have null values that would otherwise decrease the utility of the model by a considerable margin. However, it is noteworthy to mention 90% of the applications that had missing data were denied work authorization.

### Calculating Variables

In this project, we have calculated several variables using the raw variables in this dataset in combination with outside references.

**Number of applicants in a company:** We computed the number of applications submitted by a particular company and assigned the resulting number to each observation. This variable vaguely indicates the size of the company. For example: Infosys had 16,000 applications for work authorization.

**Wage:** The variable wage was given in multiple units (i.e yearly, monthly, weekly, bi weekly, hourly). In order to maintain consistency, this variable was normalized and all the wages were converted into yearly packages. However, we did find some irregularities in the units mentioned in the dataset (for ex. Hourly wage in few cases was $70,000) and we had to use our judgement to assign the right unit.

### Refining Data using different source:

The dataset included information on over 600,000 applicants and included categorical variables like State, NAICS and Job Title, which have a multitude of levels. In order to limit the number of levels we did the following treatments:

**Regions:** This variable was derived from the State variable. We used guidelines given by Bureau of Economic Analysis to consolidate the states. Using these guidelines we were able to

group the 50 states into 8 regions: New England (NE), Mideast (ME), Great Lakes (GE), Plains (PL), Southeast (SE), Rocky Mountains (RM), Farwest (FW).

**Sector:** This variable was derived using the first two codes of NAICS industry classification.

**Job Function and Domain:** The data included over 800 different job functions—in order to limit the number of levels for this variable, we decided to include only business-related domains such as Marketing, Finance, IT and Data Science, to keep this study relevant to students at Bentley. Following that change, the job-role variable was classified into 7 major categories: Analysts, Developers, QA, Database, Computer, Business and Statisticians.

## V. Initial Approach and Issues:

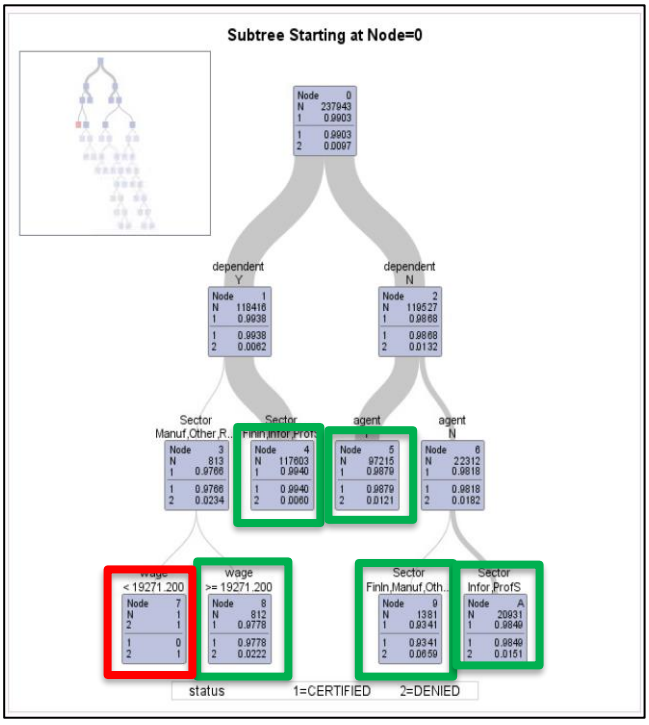
Initially the team was set to answer the question: what factors affect the likelihood of an application for work authorization get certified or denied?

However, if you refer to the table given below, we see that only 1% of the total applications for work authorization were actually denied. Thus, the data is very skewed.

**Status Summary of Applications for work authorization in FY2017**

| Status | No of Apps | Percentage |
|---|---|---|
| Certified | 235635 | 99% |
| Denied | 2309 | 1% |
| Total Applications | 237944 | |

A decision tree model was implemented to test and see if there was any predictive ability. However, given the fact that the variable 'Denied' was just observed in 1% of the dataset, each node in the decision tree was classified as Certified (with the exception of node 7, but it has only 1 observation.) **Thus, the model was not really helpful and we decided to use this dataset to answer other interesting and relevant questions.**



## VI. Primary Questions to be answered

This analysis aims to understand and explain the relationship between the predictors and our response, namely

- *"What predictors influence the likelihood that a person/company used an agent during the work authorization application?"*
- More so, do factors like industry, job function, or domain have an effect on whether or not an agency is used to for work authorization application?
- Can we predict which demographic group (socioeconomic, geographic, dependent status, etc.) is more likely to use an agent?

In the coming sections, we are going to leverage a variety of supervised and unsupervised learning techniques in an effort to answer these questions.

# VII. Supervised Learning

## A. Logistic Regression

Firstly, a logistic regression model was constructed to study the relationship between an individual's personal characteristics and whether that individual will hire an agent. Additionally, the model aims to see how well we can predict an individual's choice of hiring an agent or not.

After running multiple logistic regression models with a combination of predictors and data characteristics, we believe that an individual's *Wage*, the industry *Sector,* and the *Region* could be critical factors that affect his/her decision to hire an agent. Also, we believe there are possible interaction effects between "wage - sector" and "wage - region".

The expression of the logistic regression model is:

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

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 | . | . | . |

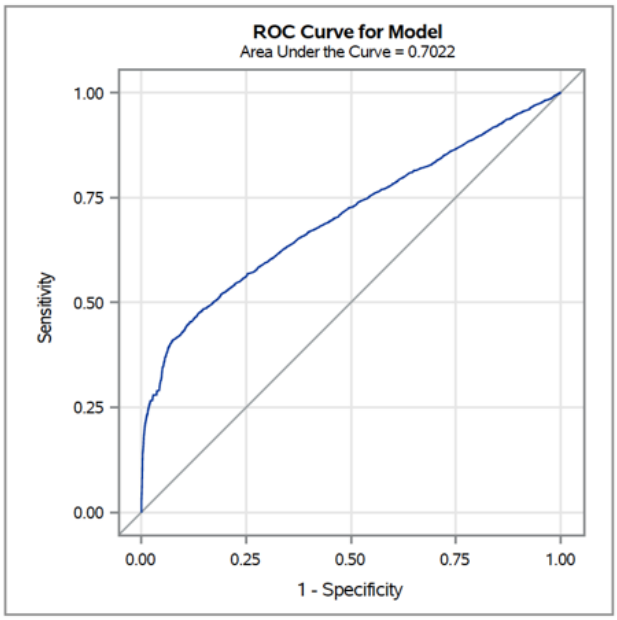
According to Maximum Likelihood estimates, the p-value for variables *"Wage", "Sector_FinIn", "Region_NE", "Wage*Region_GL", "Wage*Sector_FinIn", "Wage*Sector_Info",* and *"Wage*Sector_Manuf"* are greater than 0.05. The result indicated that these 7 variables are not statically significant, thus may not be very helpful in making predictions about agent hiring. The p-values for all other 16 variables are less than 0.05, indicating these variables are statically significant and could be used to making predictions.

## Checking Model Utility

The model's utility was then tested by constructing an ROC curve; this figure shows that AUC equals 0.7022. Both the AUC value and the ROC curve itself indicate that for a wide range of given specificity level, the model's sensitivity is higher that the case of random guessing.

Based on the classification table, the model's highest correct classification rate is at 64% at a **cutoff value equals to 0.52 or 0.54**, which means **64% of the observations are correctly classified by this model**.

The LOGISTIC Procedure




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 |
| 0.600 | 73624 | 75178 | 15717 | 73424 | 62.5 | 50.1 | 82.7 | 17.6 | 49.4 |
| 0.620 | 70135 | 78051 | 12844 | 76913 | 62.3 | 47.7 | 85.9 | 15.5 | 49.6 |
| 0.640 | 65987 | 80601 | 10294 | 81061 | 61.6 | 44.9 | 88.7 | 13.5 | 50.1 |
| 0.660 | 63289 | 81837 | 9058 | 83759 | 61.0 | 43.0 | 90.0 | 12.5 | 50.6 |
| 0.680 | 60391 | 83931 | 6964 | 86657 | 60.7 | 41.1 | 92.3 | 10.3 | 50.8 |
| 0.700 | 58372 | 84737 | 6158 | 88676 | 60.1 | 39.7 | 93.2 | 9.5 | 51.1 |
| 0.720 | 55846 | 85358 | 5537 | 91202 | 59.3 | 38.0 | 93.9 | 9.0 | 51.7 |
| 0.740 | 53046 | 85863 | 5032 | 94002 | 58.4 | 36.1 | 94.5 | 8.7 | 52.3 |
| 0.760 | 51920 | 86048 | 4847 | 95128 | 58.0 | 35.3 | 94.7 | 8.5 | 52.5 |
| 0.780 | 50463 | 86337 | 4558 | 96585 | 57.5 | 34.3 | 95.0 | 8.3 | 52.8 |
| 0.800 | 48444 | 86464 | 4431 | 98604 | 56.7 | 32.9 | 95.1 | 8.4 | 53.3 |
| 0.820 | 43439 | 86969 | 3926 | 104E3 | 54.8 | 29.5 | 95.7 | 8.3 | 54.4 |
| 0.840 | 36598 | 89321 | 1574 | 11E4 | 52.9 | 24.9 | 98.3 | 4.1 | 55.3 |
| 0.860 | 34548 | 89554 | 1341 | 113E3 | 52.2 | 23.5 | 98.5 | 3.7 | 55.7 |
| 0.880 | 33179 | 89803 | 1092 | 114E3 | 51.7 | 22.6 | 98.8 | 3.2 | 55.9 |

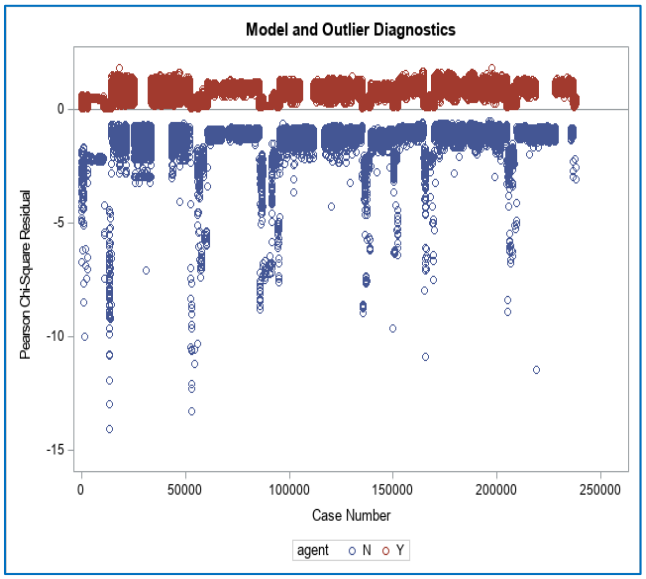
## Checking Model Validity

The model's validity was first tested by inspecting the Pearson Chi-Square Residual plot. According to the plot, we found that many residuals are involved in a pattern, indicating that the model has a dependence issue, or that there are potential relationships not revealed by or accounted for by the current model. More information is needed to build another model with additional variables.

On the other hand, an HL test was conducted to further check model validity. According to Figure 3, the p-value for the HL test was less than 0.0001, ***so we reject the null hypothesis that the logistic model used accurately describes the data.***

This result indicates that the model's ability to make predictions is questionable, and that this regression model is not a good way to study our question. Based on this conclusion, the next steps of our study do not apply to this logistic regression model and we will not further explain nor use this model to make any predictions.

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




## Limitations

The results from checking model validity indicate that the logistic regression model we built does not accurately describe the data, despite the fact that AUC value and ROC curve indicate a positive prediction ability, and despite the fact that 16 out of 23 coefficients are statistically significant. A similar issue was noted in other research papers. According to Kramer, in his study on patient mortality with logistic regression, he noted that "the Hosmer-Lemeshow test was sensitive to sample size"[1]. Kramer's study concluded that the percentage of the HL tests he did that were significant at p-value < 0.05 significantly increased when the sample size he used increased from 5,000 to 50,000. Since we have more than 240,000 observations in our study, there is a high possibility that our logistic regression model passes the HL test. However, a

[1] 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

significant Hosmer-Lemeshow test does not necessarily mean that a predictive model is not useful or suspect. In future studies, the validity of a logistic regression model could be tested using other methods, such as Pearson Chi-Squared test or Deviance Chi-Squared test.

In addition, since the residuals of the current model show a pattern, we believe that additional background information should be obtained by conducting interviews or focus groups prior to the beginning of any further study. More variables or interaction terms may need to be added into new models.

## B. Classification Tree

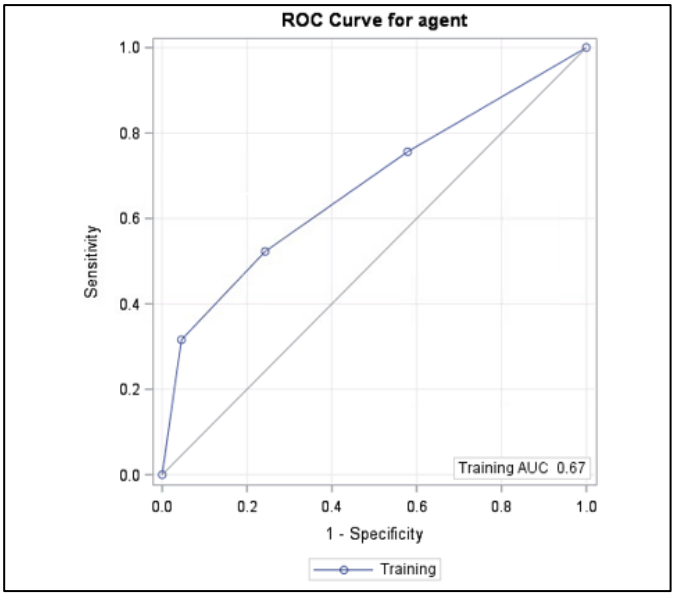
### Classification Tree #1

We first examine the relationship between whether the applicant has an agent or not and the predictors included region and sector. The model fitting is conducted using SAS and the full output can be found in the Appendix.

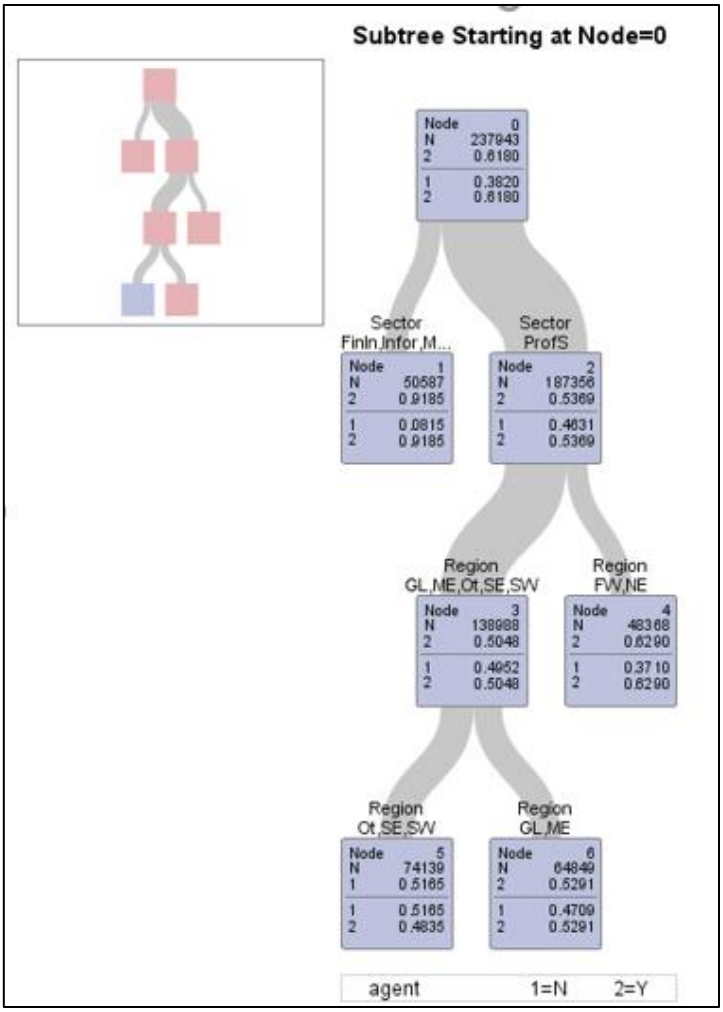
The fitted decision tree model demonstrates a modest predictive performance as indicated by the area under the receiver operating characteristics curve of 0.67 reported in below chart. From this model, we are able to classify whether or not an applicant will choose an agent based on the region and sector the applicant belongs to.

Sector shows a variable importance of 107.7 and region shows a variable importance of 35.37 from the variable importance table. Indicating that Sector is relatively more important than the region in determining if a visa agency is hired.

Receiver Operating Curve for Classification tree #1



Classification Tree #1 with ‘region’ and ‘sector’ as Predictors.




The above figure shows the decision tree with sector and region as predictors. There are some insights we can draw from this tree.

- According to the model, applicants from sectors other than Professional, Scientific, and Technical Services are classified as likely to hire an agency to apply for their H-1B. The probability for these individuals to use an agent is roughly 92%. (Found in Leaf node 1)
- For applicants from Professional, Scientific, and Technical Services, if they are from the regions Far West and New England, the model classifies them to be likely to hire an agency. (Found in Leaf node 4). If they are from regions including Southeast, Southwest and Other, then they are classified as individuals who are likely to apply for the H-1B visa by themselves. (Indicated in Leaf node 5)

The misclassification rate for this model is 37.17% which means that **this model classifies about 63% of the observations accurately**. The **sensitivity is 75.62%** as reported in the confusion matrix from Table 1. The specificity is only 42.13%. Thus, our model does a much better job when predicting event (agent=’Y’) vs. non-event (agent=’N’).

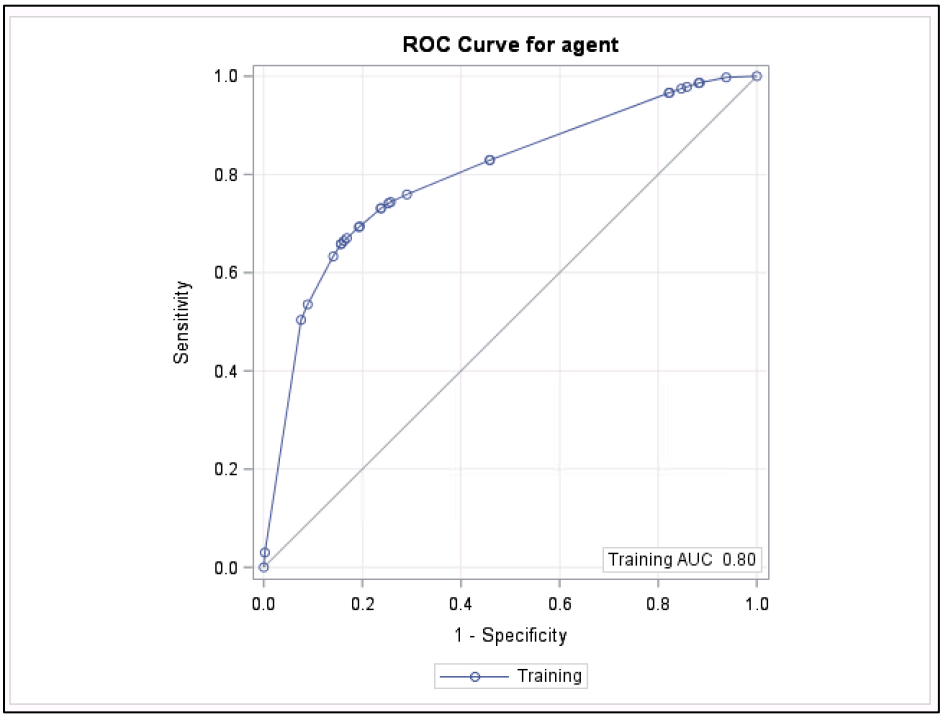
Table 1. Confusion Matrix for Decision Tree 1.

| Model-Based Confusion Matrix | | | |
|---|---|---|---|
| | Predicted | | Error Rate |
| Actual | N | Y | |
| 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 |

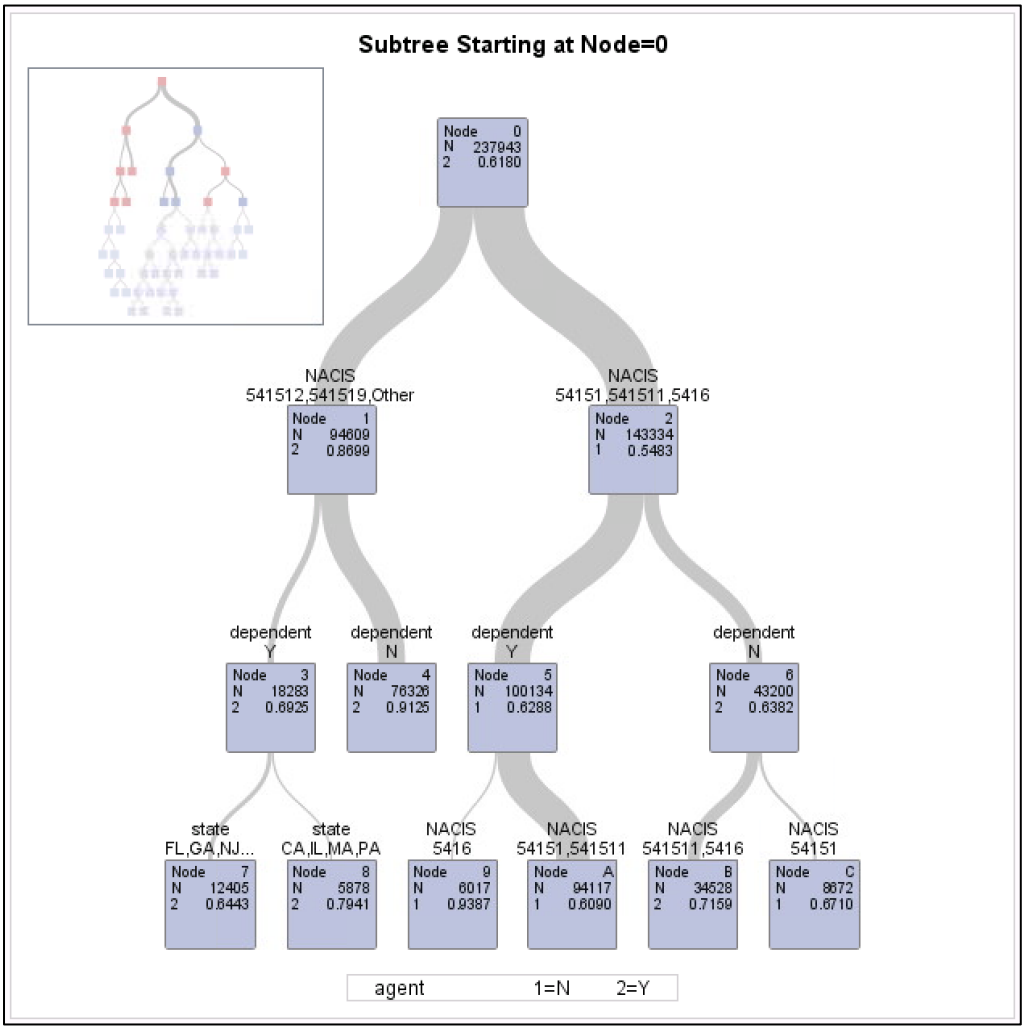
## Classification Tree #2

To systematically study the relationship between whether the applicant has an agent or not with predictors provided from the dataset, we built a second decision tree to see if other variables can do a better job of predicting our response variable. In the second tree, we used variables like NAICS, State and Dependent.




Receiver Operating Characteristic Curve for Decision Tree 2.

The fitted decision tree model demonstrates a good predictive performance as indicated by the area under the receiver operating characteristics curve of 0.80. From this model, we find that there is evidence that NAICS code (North American Industry Classification System) is a more important predictor in classifying whether the applicant has an agent or not as opposed to the 'dependent' and 'state' variables, as NAICS has the highest variable of importance (154.7) from the variable importance table (Appendix).




Classification Tree #2 with 'State', 'dependent' and 'NAICS' as Predictors

The above figure shows classification tree with State, dependent and NAICS as predictors. There are some insights we can draw from this tree:

- According to the classification tree, Applicants from industries with NAICS code 541512, 541519 and other NAICS code with minor number of applicants are classified as individuals who are likely to hire an agent with a probability of 87%.

  Among these applicants, if they do not have a dependent then they are classified as applicants who hire an agent. If they have a dependent, then only applicants from California, Illinois, Massachusetts, and Pennsylvania are classified as applicants who would be getting an agent to help them apply for the H-1B visa.

- In contrast, the model classified applicants from industries with NAICS code 54151, 541511 and 5416 as individual who would apply H-1B by themselves with a probability of 55%. If they have a dependent, they are classified as individual who would apply for the visa by themselves with a probability of 63%. If they do not have a dependent, they are classified as applicants who would apply the H-1B with an agent. In addition, if the applicant has a dependent and belongs to industry with NAICS code 5416, then they are classified as applicants who will apply with an agent.

The misclassification rate for the final tree is reported to be 25.64% from the Model-Based Fit Statistics for Selected Tree table (Table 2). Thus, **the tree misclassifies about 26% of the observations in the original dataset**. We can also see that the sensitivity of the final **model is 74.12% and the specificity is 73.74%.** The model has almost equal capability to predict the event (agent 'Y') vs. non-event (agent 'N').

Table 2. Confusion Matrix for Decision Tree 2.

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 |

**Model Limitations:**

The classification tree model generally provides good predictions on whether the applicants apply for H-1B visa with an agent or not. However, there are also some limitations to the decision trees.

- We are not able to check model validity for our model due to the nature of decision trees
- they don't have any statistical basis nor do they make any assumptions. No assumptions can give us more flexibility when building the model, but it can also give us a misleading model.
- In addition, decision trees are highly dependent upon the dataset on which they were created and are prone to overfitting. Thus, if we use H-1B dataset from another year, it will potentially give different results.

In future studies, we can take a different approach and employ more robust techniques such as random forest. Random forest aggregates many decision trees to limit overfitting as well as errors due to bias and therefore yields more useful results[2].

# VIII. Unsupervised Learning

## Cluster Analysis

### Cluster Development

Our dataset has 3 quantitative variables namely: Yearly Wage of the employees, Number of Applications a given company has submitted for work authorization in FY2017 and the proposed duration of the employment. Please note that here, the variable number of applications a given company has submitted is used as a proxy for company size (i.e number of employees in the company)

In order to understand profiles of the applicants for the H-1B process, we conduct a hierarchical cluster analysis using average linkage, median and complete linkage in SAS. Please note that a dendrogram could not be generated due to high number of observations. Due to this, multiple combinations of clusters were executed in order to get a clustering that would satisfactorily represent our dataset. The outputs of means procedure for 3 and 4 clustering for average, median and complete linkage procedures are given in the Appendix.

After testing out for the best number of clusters to define, we find that the customers can be grouped into **4 clusters** to satisfactorily represent our dataset using **complete linkage method.** Since, we do not have any additional quantitative variable to verify the if the means of the clusters are truly different, we take a different approach instead. We compare the clusters based on the regions and roles to visually verify if there is a difference.

**The MEANS Procedure**

| CLUSTER | N Obs | Variable | Label | N | Mean | Std Dev | Minimum | Maximum |
|---|---|---|---|---|---|---|---|---|
| 1 | 467 | duration | duration | 467 | 1069.51 | 123.7439445 | 34.0000000 | 1096.00 |
| | | wage | wage | 467 | 89603.58 | 12256.29 | 72821.00 | 118997.00 |
| | | Companyapps | Companyapps | 467 | 2099.87 | 4397.10 | 1.0000000 | 16141.00 |
| 2 | 57 | duration | duration | 57 | 1094.86 | 0.6665100 | 1093.00 | 1096.00 |
| | | wage | wage | 57 | 135780.93 | 14254.25 | 120662.00 | 169749.00 |
| | | Companyapps | Companyapps | 57 | 855.7894737 | 2273.89 | 1.0000000 | 16141.00 |
| 3 | 550 | duration | duration | 550 | 1062.90 | 128.0897674 | 61.0000000 | 1096.00 |
| | | wage | wage | 550 | 60315.98 | 8513.84 | 35776.00 | 74797.00 |
| | | Companyapps | Companyapps | 550 | 1748.33 | 3732.36 | 1.0000000 | 16141.00 |
| 4 | 4 | duration | duration | 4 | 1094.75 | 0.5000000 | 1094.00 | 1095.00 |
| | | wage | wage | 4 | 195837.00 | 13342.74 | 184288.00 | 214947.00 |
| | | Companyapps | Companyapps | 4 | 2156.50 | 3418.72 | 128.0000000 | 7273.00 |

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

**Clusters vs Role and Region of applicants**

| 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% |

| Cluster | Region | |
|---|---|---|
| 1 | FW | 29.34% |
| | GL | 9.85% |
| | ME | 22.27% |
| | NE | 7.71% |
| | Other | 6.42% |
| | SE | 13.28% |
| | SW | 11.13% |
| 2 | FW | 77.19% |
| | ME | 17.54% |
| | NE | 1.75% |
| | SW | 3.51% |
| 3 | FW | 11.09% |
| | GL | 17.09% |
| | ME | 21.09% |
| | NE | 5.45% |
| | Other | 7.45% |
| | SE | 22.18% |
| | SW | 15.64% |
| 4 | FW | 75.00% |
| | ME | 25.00% |

The clusters are as follows:
The first cluster can be labeled as **Corporate Specialists**. This cluster constitutes of employees with a wide mix of roles mainly split between Analysts and Developers with a median salary of about 90k. These employees are found in mainly found in the mainly in regions with big opportunities like Middle East (New York, New Jersey, etc) and Far West (California, Washington, Nevada).

The second cluster can be labeled as **High skilled techies**. This cluster constitutes of employees with high median salary of above 140k. These applicants are mainly focused in the California and Washington region with developer being the highest frequency role (75%).

The third cluster can be labeled as **Corporate Associates**. This cluster is similar to cluster 1 except for the fact that they have a lower median salary of about $50k to $60k. These are probably applicants who have newly started their career.

The fourth cluster can be labeled as **high-level bosses**. The applicants in this cluster has astonishingly high yearly salaries almost close to $200k per year. These applicants are mainly on the director or VP level in the computer architecture domain. However, the sample size is very small for cluster 4 but given the variation in the wage, these observations deserve to be their own cluster.

**Limitations**

It is important to note that the Cluster Analysis procedure does not have any statistical basis. Thus, to conclude the right number of cluster groupings or specifying the right clustering is subjective and requires domain knowledge. Additionally, given the large size of this dataset, the university version of SAS could not support the computation of the large number of observations our data contained. Thus, in order to support the processing power for the student edition of SAS,

we had to take samples of the original dataset for clustering. However, it is worthy to note that multiple samples of equal sizes were taken and in every iteration the clustering was more or less very similar across these samples. However, the clustering is based on smaller samples and may still be dependent on the data. Thus, we cannot generalize these findings to a large population with complete certainty.

## IX. Summary & Concluding Remarks

With the increase in immigrant population in USA, the applications for H1B and other work-related authorizations have increased over the years. Visa application agencies have tremendous market opportunity with an average of 600,000 applicants applying for work related visa every year in USA alone. This study was narrowed down to keep it relevant to students at Bentley by limiting the number of distinct job roles for this study.

Multiple supervised learning techniques were used to help understand what factors can determine whether an applicant takes aid of such visa application agents. With logistic regression, we were able to determine that sector and the region of the applicant are important factors. However, since model validity was in question, we carried out classification tree. The findings of both the supervised learning techniques concur that the sector in which the applicant belongs to is a strong indicator as compared to other parameters. The proposed models even with their share of limitations can help visa application agencies with a starting point to narrow down their prospects.

Lastly with the help of cluster analysis techniques, we were able to group the applications into 4 different profiles. As already discussed, there are quite a few limitations in the model mainly due to the inherent properties of Cluster analysis and Classification trees. In the future, in order to make the study more robust, data from different fiscal years can be compared and the performance of the models can be compared for these time periods to validate if these predictors indeed have predictive capabilities in determining whether or not an employee takes help of a visa applications agency.

## X. References

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

## XI. Data Sources

- **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/**

- **US Department of Labor**
  **https://www.dol.gov/**

# XII. Appendix

**Variables in the final dataset**

The final list of variables used for analysis after applying transformations in the dataset are given below

| Sr. No | Variable Name | Characteristics |
|---|---|---|
| 1 | CASE_NUMBER | This is a unique categorical identifier tied to each applicant |
| 2 | Case status | This variable is either “CERTIFIED” or “DENIED” and describes whether an applicant was granted work authorization |
| 3 | Employment Duration | This variable was calculated using the employment start date and employment end date on the application. This is a numerical variable with the unit mentioned in days |
| 4 | Number of Applicants in a company | This numerical that describes the number of applications for work authorization a given company has submitted in FY2017. |
| 5 | wage | Wage is a numeric variable calculated using two variables from the original data set ‘Wage’ and ‘Unit’. For data analysis, all the salaries were normalized to yearly wage |
| 6 | domain | This is a categorical variable that describes the primary area of operation under which the individual applicant will work. This variable was derived from the variable ‘job-title’ to include only Marketing, Finance, IT and Data Science to keep this study relevant to students at Bentley. |
| 7 | role | This is a categorical variable that was derived from the variable ‘job title’ from the original data set. Since job title had more than 800 levels, the new variable ‘role’ was created to group similar job-titles and to focus only the job titles that were relevant for this study. |
| 8 | Sector | This is a categorical, describes the primary business operation of the employer. This variable was derived from the NAICS industry classification mentioned in the dataset. |
| 9 | fulltime | This is a categorical that indicates whether an employee is full-time (“1”) or not (“0”). |
| 10 | dependent | This is a categorical describes whether the applicant claims a dependent on their application. This is a dummy variable, where “1” indicates that a dependent has been claimed, and “0” indicates no dependent. |

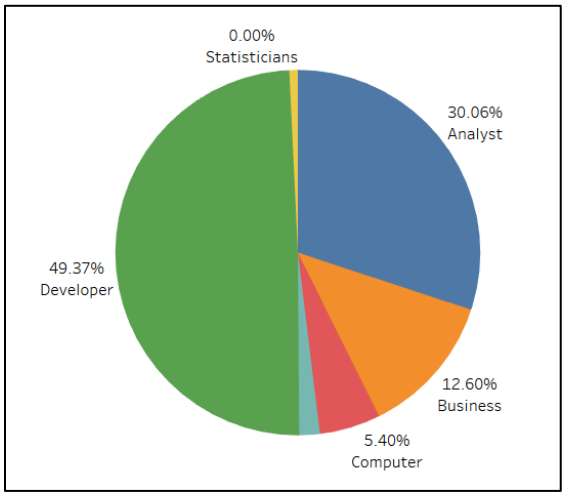
| | | |
|---|---|---|
| 12 | Region | Region is a categorical variable, that was derived using the variable 'State'. Multiple states were grouped under one region using classification by the Bureau of Economic Analysis. For example, states like MA, RH, NH would come under the broad category of 'New England Region'. |
| 13 | agent | This categorical variable and also considered as our dependent variable in logistic regression and decision tree. This variable is a dummy variable, where "1" indicates that an applicant/company used an agent, and "0" indicates that they did not. |
| 14 | Count of Blanks | This variable was calculated to understand how many fields in the application were left blank. |
| 15 | NAICS | This is a categorical variable. NAICS codes are industry classifications used to classify and describe the operations of companies. Two-digit codes define the broader industry, while 4 and 5+ digit codes are more granular. In this analysis, multiple industries were grouped together under the category 'other' and the industries focused were included:<br>o NAICS 54151 - Computer Systems Design and Related Services<br>o NAICS 541511 - Custom Computer Programming Services<br>o NAICS 541512 - Computer Systems Design Services<br>o NAICS 541519 - Other Computer Related Services<br>o NAICS 5416 - Management, Scientific, and Technical Consulting Services |
| 16 | state | This is a categorical variable describing the U.S. State in which an applicant would reside during the duration of their granted stay. |

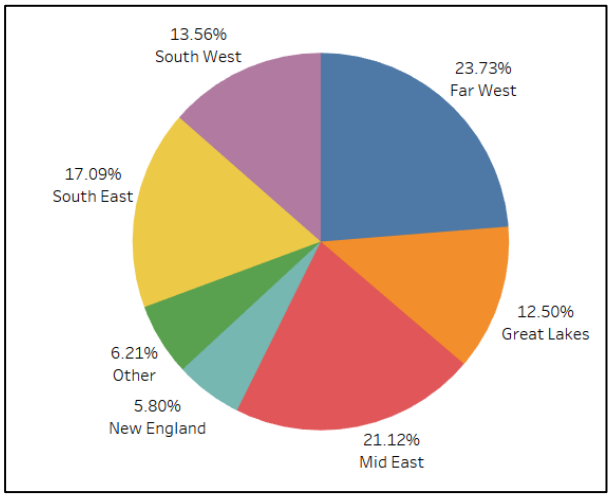
## Descriptive Statistics and Visual Data Summary

**Distribution of Applicants choosing to take an agent**

| Agent? | # of Records | Percentage |
|---|---|---|
| N | 90896 | 38% |
| Y | 147048 | 62% |
| Grand Total | 237944 | |

**Distribution of job roles and regions**




**Distribution of Sector**

| 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 | |

Since the Sector Professional, Scientific, and Technical Services constitute to a major part of the population, we also considered NAICS as an indicator for industry with focus on this sector. Where ProfScienTech (NAICS 54 series) was further divided into following:

- NAICS 54151 - Computer Systems Design and Related Services
- NAICS 541511 - Custom Computer Programming Services
- NAICS 541512 - Computer Systems Design Services
- NAICS 541519 - Other Computer Related Services
- NAICS 5416 - Management, Scientific, and Technical Consulting Services

**Distribution of Quantitative variables**

| | Mean | Max | Min | Std Dev |
|---|---|---|---|---|
| Wage ($/year) | 76931 | 434200 | 15080 | 22487.23 |
| Duration (days) | 1062 | 1096 | 1 | 163.9643 |
| No of Applicants from the company | 1771 | 16141 | 1 | 3863.082 |

## Decision Tree with sector & region as predictors

**The SAS System**

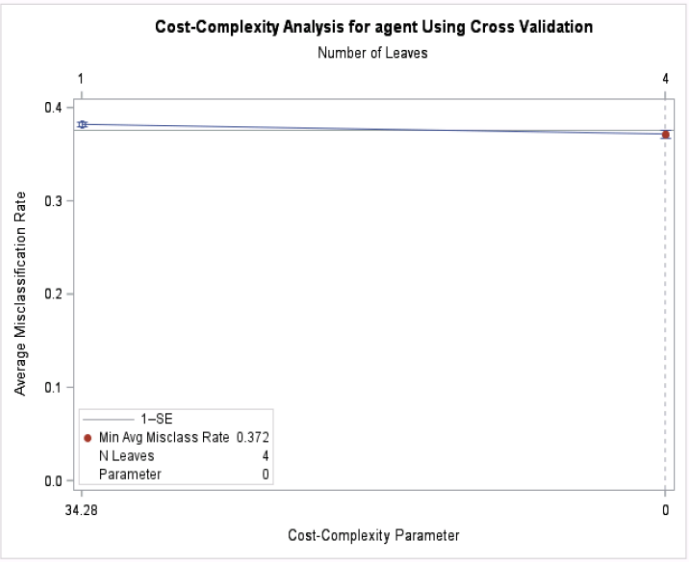
**The HPSPLIT Procedure**

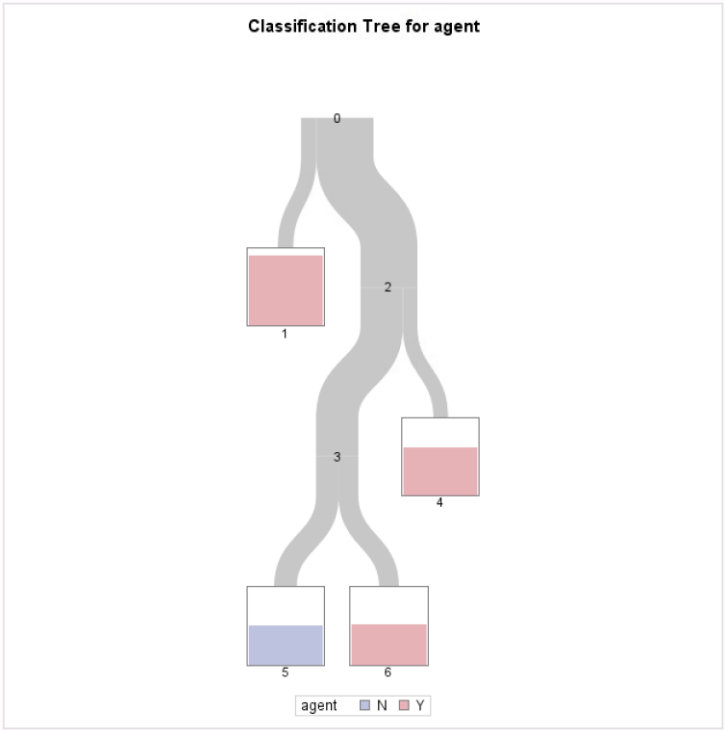
| Performance Information | |
|---|---|
| Execution Mode | Single-Machine |
| Number of Threads | 4 |

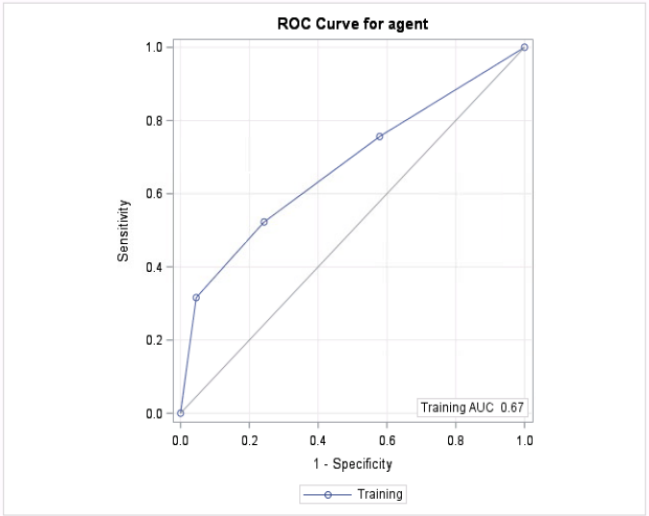
| Data Access Information | | | |
|---|---|---|---|
| Data | Engine | Role | Path |
| WORK.DATA | V9 | Input | On Client |

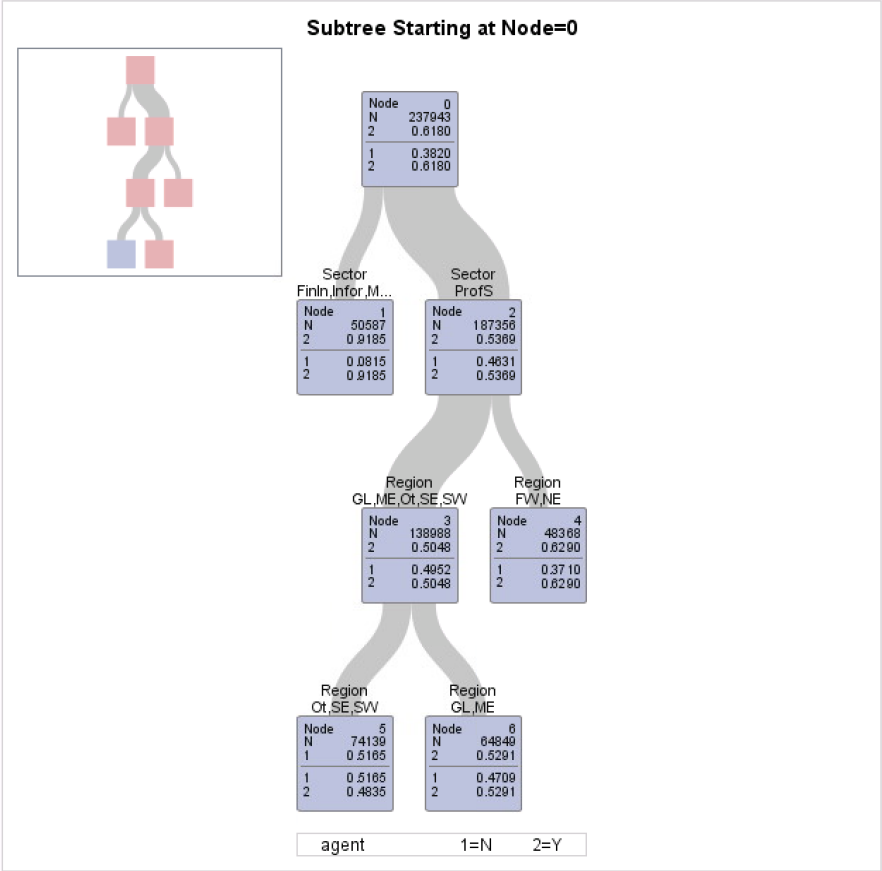
| Model Information | |
|---|---|
| Split Criterion Used | Entropy |
| Pruning Method | Cost-Complexity |
| Subtree Evaluation Criterion | Cost-Complexity |
| Number of Branches | 2 |
| Maximum Tree Depth Requested | 10 |
| Maximum Tree Depth Achieved | 8 |
| Tree Depth | 3 |
| Number of Leaves Before Pruning | 42 |
| Number of Leaves After Pruning | 4 |
| Model Event Level | Y |

| | |
|---|---|
| Number of Observations Read | 237944 |
| Number of Observations Used | 237943 |






| Variable Importance | | | |
|---|---|---|---|
| | Training | | |
| Variable | Relative | Importance | Count |
| Sector | 1.0000 | 107.7 | 1 |
| Region | 0.3283 | 35.3704 | 2 |

**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 with NAICS, dependent & state as predictors

**The SAS System**

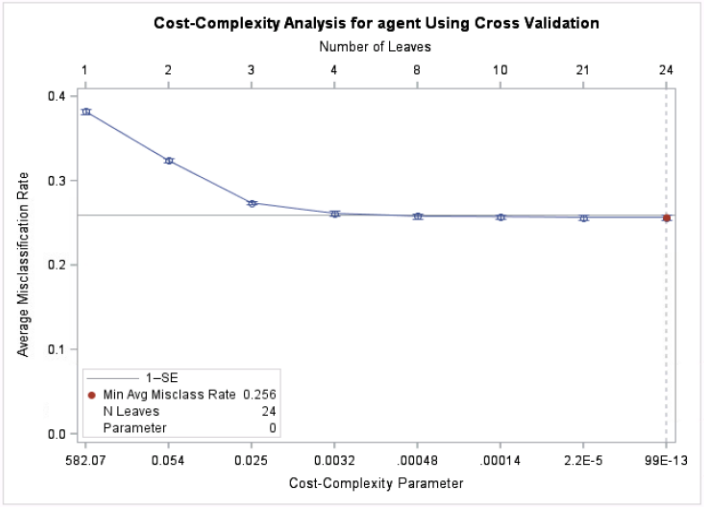
**The HPSPLIT Procedure**

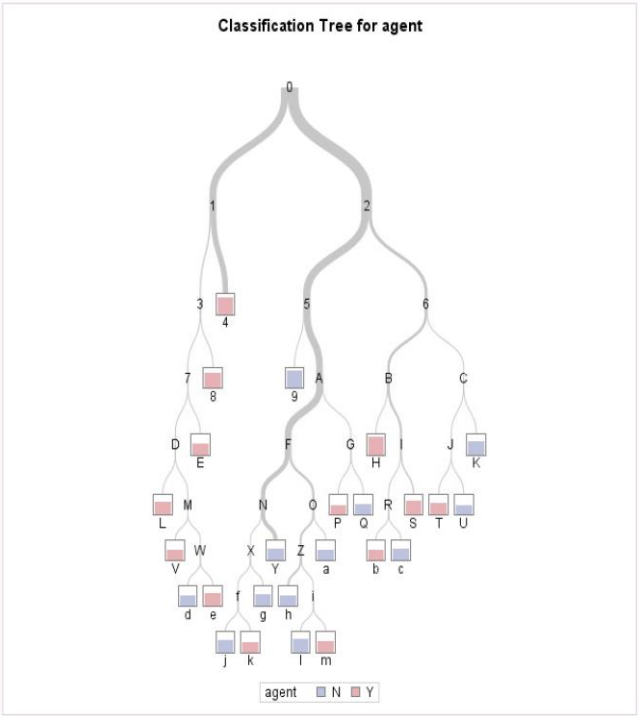
| Performance Information | |
|---|---|
| Execution Mode | Single-Machine |
| Number of Threads | 4 |

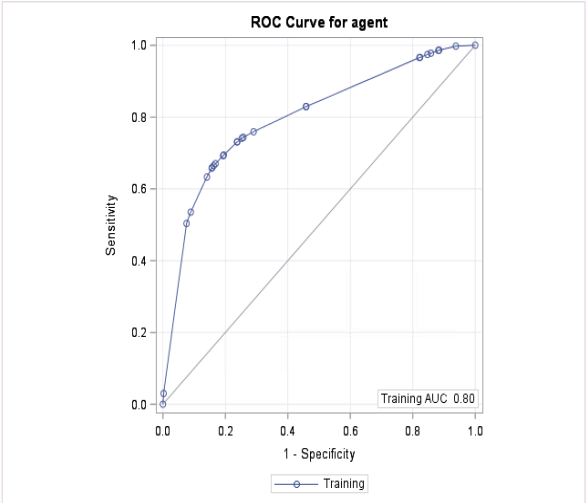
| Data Access Information | | | |
|---|---|---|---|
| Data | Engine | Role | Path |
| WORK.H1B | V9 | Input | On Client |

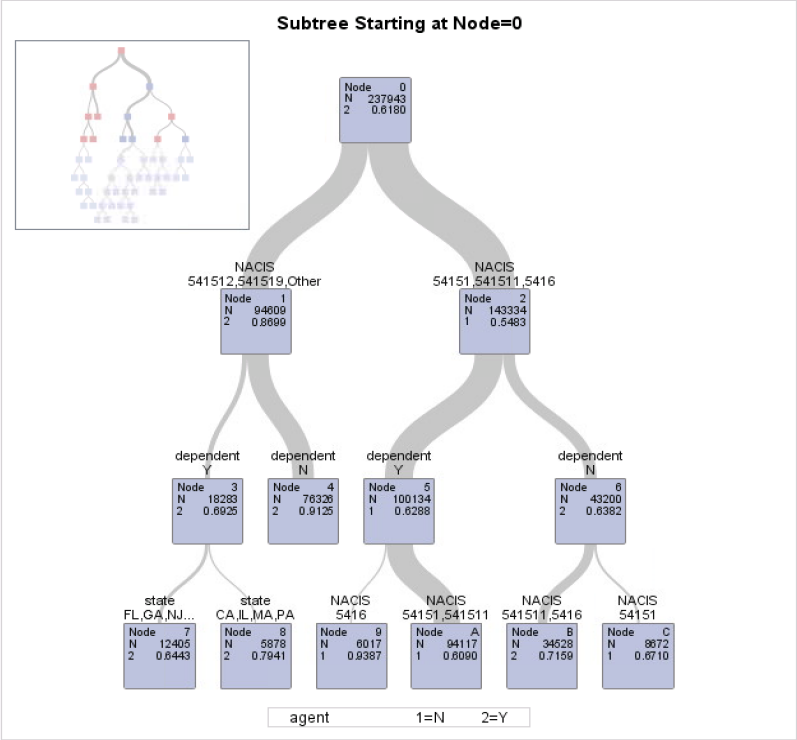
| Model Information | |
|---|---|
| Split Criterion Used | Entropy |
| Pruning Method | Cost-Complexity |
| Subtree Evaluation Criterion | Cost-Complexity |
| Number of Branches | 2 |
| Maximum Tree Depth Requested | 10 |
| Maximum Tree Depth Achieved | 9 |
| Tree Depth | 8 |
| Number of Leaves Before Pruning | 120 |
| Number of Leaves After Pruning | 25 |
| Model Event Level | Y |

| | |
|---|---|
| Number of Observations Read | 237944 |
| Number of Observations Used | 237943 |






| Variable Importance | | | |
|---|---|---|---|
| | Training | | |
| Variable | Relative | Importance | Count |
| NACIS | 1.0000 | 154.7 | 9 |
| dependent | 0.4891 | 75.6918 | 2 |
| state | 0.2537 | 39.2510 | 13 |

| Model-Based Confusion Matrix | | | |
|---|---|---|---|
| | Predicted | | |
| Actual | N | 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 |

## CLUSTERING

Means Procedure for Samples of Main Dataset

| Variable | Label | N | Mean | Std Dev | Minimum | Maximum |
|---|---|---|---|---|---|---|
| duration | duration | 1078 | 1067.57 | 122.65 | 34.00 | 1096.00 |
| wage | wage | 1078 | 77496.7 | 23645.41 | 35776.00 | 214947.00 |
| Companyapps | Companyapps | 1078 | 1854.94 | 3981.60 | 1.00 | 16141.00 |

**The MEANS Procedure**

Below are the outputs for Clustering using different linkage methods.

In all of the methods below except complete linkage, the clustering was not satisfactory as we would always be left out with a cluster with 'too few' observations to make any kind of generalization.

- **4 Clusters using Average Linkage Clustering**

The MEANS Procedure

| CLUSTER | N Obs | Variable | Label | N | Mean | Std Dev | Minimum | Maximum |
|---|---|---|---|---|---|---|---|---|
| 1 | 763 | duration | duration | 763 | 1061.99 | 131.1554744 | 61.0000000 | 1096.00 |
| | | wage | wage | 763 | 65411.71 | 11130.84 | 35776.00 | 85883.00 |
| | | Companyapps | Companyapps | 763 | 1947.99 | 4159.71 | 1.0000000 | 16141.00 |
| 2 | 279 | duration | duration | 279 | 1079.29 | 103.9756457 | 34.0000000 | 1096.00 |
| | | wage | wage | 279 | 101103.73 | 10939.53 | 84094.00 | 128107.00 |
| | | Companyapps | Companyapps | 279 | 1766.36 | 3677.57 | 1.0000000 | 16141.00 |
| 3 | 35 | duration | duration | 35 | 1094.94 | 0.5912528 | 1093.00 | 1096.00 |
| | | wage | wage | 35 | 148842.78 | 17580.43 | 130603.00 | 193918.00 |
| | | Companyapps | Companyapps | 35 | 567.1428571 | 1335.92 | 2.0000000 | 7273.00 |
| 4 | 1 | duration | duration | 1 | 1095.00 | . | 1095.00 | 1095.00 |
| | | wage | wage | 1 | 214947.00 | . | 214947.00 | 214947.00 |
| | | Companyapps | Companyapps | 1 | 642.0000000 | . | 642.0000000 | 642.0000000 |

- **3 Clusters using Average Linkage Clustering**

The MEANS Procedure

| CLUSTER | N Obs | Variable | Label | N | Mean | Std Dev | Minimum | Maximum |
|---|---|---|---|---|---|---|---|---|
| 1 | 1042 | duration | duration | 1042 | 1066.62 | 124.6487333 | 34.0000000 | 1096.00 |
| | | wage | wage | 1042 | 74968.40 | 19304.31 | 35776.00 | 128107.00 |
| | | Companyapps | Companyapps | 1042 | 1899.36 | 4035.34 | 1.0000000 | 16141.00 |
| 2 | 35 | duration | duration | 35 | 1094.94 | 0.5912528 | 1093.00 | 1096.00 |
| | | wage | wage | 35 | 148842.78 | 17580.43 | 130603.00 | 193918.00 |
| | | Companyapps | Companyapps | 35 | 567.1428571 | 1335.92 | 2.0000000 | 7273.00 |
| 3 | 1 | duration | duration | 1 | 1095.00 | . | 1095.00 | 1095.00 |
| | | wage | wage | 1 | 214947.00 | . | 214947.00 | 214947.00 |
| | | Companyapps | Companyapps | 1 | 642.0000000 | . | 642.0000000 | 642.0000000 |

- **3 Clusters using Median Linkage Clustering**

The MEANS Procedure

| CLUSTER | N Obs | Variable | Label | N | Mean | Std Dev | Minimum | Maximum |
|---|---|---|---|---|---|---|---|---|
| 1 | 893 | duration | duration | 893 | 1064.20 | 127.5834253 | 61.0000000 | 1096.00 |
| | | wage | wage | 893 | 69284.23 | 13981.75 | 35776.00 | 97841.00 |
| | | Companyapps | Companyapps | 893 | 1863.57 | 3946.12 | 1.0000000 | 16141.00 |
| 2 | 181 | duration | duration | 181 | 1083.60 | 94.9594521 | 34.0000000 | 1096.00 |
| | | wage | wage | 181 | 115399.75 | 16827.21 | 94141.00 | 169749.00 |
| | | Companyapps | Companyapps | 181 | 1805.67 | 4181.84 | 1.0000000 | 16141.00 |
| 3 | 4 | duration | duration | 4 | 1094.75 | 0.5000000 | 1094.00 | 1095.00 |
| | | wage | wage | 4 | 195837.00 | 13342.74 | 184288.00 | 214947.00 |
| | | Companyapps | Companyapps | 4 | 2156.50 | 3418.72 | 128.0000000 | 7273.00 |

- **4 Clusters using Median Linkage Clustering**

The MEANS Procedure

| CLUSTER | N Obs | Variable | Label | N | Mean | Std Dev | Minimum | Maximum |
|---|---|---|---|---|---|---|---|---|
| 1 | 893 | duration | duration | 893 | 1064.20 | 127.5834253 | 61.0000000 | 1096.00 |
| | | wage | wage | 893 | 69284.23 | 13981.75 | 35776.00 | 97841.00 |
| | | Companyapps | Companyapps | 893 | 1863.57 | 3946.12 | 1.0000000 | 16141.00 |
| 2 | 172 | duration | duration | 172 | 1082.99 | 97.3886227 | 34.0000000 | 1096.00 |
| | | wage | wage | 172 | 112928.21 | 13130.01 | 94141.00 | 147846.00 |
| | | Companyapps | Companyapps | 172 | 1883.16 | 4275.62 | 1.0000000 | 16141.00 |
| 3 | 9 | duration | duration | 9 | 1095.11 | 0.3333333 | 1095.00 | 1096.00 |
| | | wage | wage | 9 | 162633.56 | 6689.78 | 154003.00 | 169749.00 |
| | | Companyapps | Companyapps | 9 | 324.7777778 | 354.4325669 | 5.0000000 | 873.0000000 |
| 4 | 4 | duration | duration | 4 | 1094.75 | 0.5000000 | 1094.00 | 1095.00 |
| | | wage | wage | 4 | 195837.00 | 13342.74 | 184288.00 | 214947.00 |
| | | Companyapps | Companyapps | 4 | 2156.50 | 3418.72 | 128.0000000 | 7273.00 |

- **4 Clusters using Complete Linkage clustering**

The MEANS Procedure

| CLUSTER | N Obs | Variable | Label | N | Mean | Std Dev | Minimum | Maximum |
|---|---|---|---|---|---|---|---|---|
| 1 | 467 | duration | duration | 467 | 1069.51 | 123.7439445 | 34.0000000 | 1096.00 |
| | | wage | wage | 467 | 89603.58 | 12256.29 | 72821.00 | 118997.00 |
| | | Companyapps | Companyapps | 467 | 2099.87 | 4397.10 | 1.0000000 | 16141.00 |
| 2 | 57 | duration | duration | 57 | 1094.86 | 0.6665100 | 1093.00 | 1096.00 |
| | | wage | wage | 57 | 135780.93 | 14254.25 | 120662.00 | 169749.00 |
| | | Companyapps | Companyapps | 57 | 855.7894737 | 2273.89 | 1.0000000 | 16141.00 |
| 3 | 550 | duration | duration | 550 | 1062.90 | 128.0897674 | 61.0000000 | 1096.00 |
| | | wage | wage | 550 | 60315.98 | 8513.84 | 35776.00 | 74797.00 |
| | | Companyapps | Companyapps | 550 | 1748.33 | 3732.36 | 1.0000000 | 16141.00 |
| 4 | 4 | duration | duration | 4 | 1094.75 | 0.5000000 | 1094.00 | 1095.00 |
| | | wage | wage | 4 | 195837.00 | 13342.74 | 184288.00 | 214947.00 |
| | | Companyapps | Companyapps | 4 | 2156.50 | 3418.72 | 128.0000000 | 7273.00 |