2022

Team Galaxy
(Prashant Raghuwanshi & Jay Pfister)
DSC630-T301 Predictive Analytics (2223-1)

Executive Summary

Every year, the United States Bureau of Labor Statistics conducts the American Time Use Survey (ATUS). The purpose of this survey is to determine the amount of time people in the United States spend on different things. The survey has questions pertaining to topics ranging from employment, to the number of dependents a respondent has, to the amount of time they spend with those dependents, to the amount of time they spend watching TV.

The bureau uses this survey to collect information on what it is the public is spending their time on and determine what impacts that may be having on the labor force and why. For example, a number of the questions pertain to individual employment. Is the respondent employed? If so, in what industry? If not, why not? Are they injured or retired or just temporarily out of work? This information provides a snapshot of the life of the average American's time and how they spend it.

This could speak to a number of things, such as the amount of work the average American needs to perform to provide for their family, however large, or as a contrast between that work and the amount of free time they have for themselves and how they choose to use it.

This project aims to take the collected data from the ATUS survey and determine which of these factors best identifies which of these factors indicates that a respondent may be in need of financial services. The theory being that a gainfully employed individual supporting themselves or their family will likely need access to financial services provided by banking institutions, such as savings accounts and certificates of deposit (CDs).

This information can be provided to those institutions with the intent of helping them to better target those individuals who may be in need of those services with marketing efforts in order to better serve both the individual and the institution itself.

Abstract

This term-end project aims to evaluate the students' ability to identify a business problem to address through predictive analytics. The goal is to select appropriate models and model specifications and apply the respective methods to enhance data-driven decision-making related to the business problem.

We identified the potential use of predictive analytics, formulate the problem, identify the right sources of data, analyze data and prescribe actions to improve not only the process of decision making but also the outcome of decisions.

We are targeting a sub set of survey respondents who we have identified as gainfully and regularly employed and potentially in need of financial services such as savings accounts or CDs.

The Data

This data set contains case-specific variables collected in ATUS (that is, variables for which there is one value for each respondent). These include, for example, labor force and earnings information, total time providing secondary childcare, and ATUS statistical weights. There is one record for each ATUS respondent.

Below is a simplified example. The variable TUCASEID identifies each household, and the variable TULINENO identifies each individual within the household. The example contains responses from five individuals; note that the respondent always has TULINENO = 1. In the example, each respondent has a corresponding statistical weight for use in generating estimates representative of the U.S. civilian, noninstitutionalized population (TUFINLWGT), and values for school enrollment (TESCHENR), labor force

status (TELFS), and total number of minutes spent alone on the diary day (TRTALONE). The actual ATUS Respondent file contains more variables and records.

TUCASEID	TULINENO	TUFINLWGT	TESCHENR	TELFS	TRTALONE
20060101020210	1	22261358.19	1	1	40
20060101020211	1	5019645.31	1	1	350
20060101020212	1	2926068.74	1	5	0
20060101020213	1	25780574.07	2	5	556
20060101020214	1	3414645.94	1	4	100

	TEERNPER_1.0	TEERNPER_6.0	TEERNPER_2.0	TEERNPER_3.0	TEERNPER_5.0	TEHRUSLT_25.0	TEHRUSLT_40.0	TEHRUSLT_50.0	TEHRUSLT_45.0	TEHRU
TEERNPER_1.0	1.000000	0.735187	0.295063	0.292895	0.191691	0.576347	0.325572	0.189185	0.149678	
TEERNPER_6.0	0.735187	1.000000	0.117989	0.117111	0.075984	0.427135	0.254514	0.182306	0.149754	
TEERNPER_2.0	0.295063	0.117989	1.000000	0.045137	0.027518	0.163306	0.065357	0.042452	0.027255	
TEERNPER_3.0	0.292895	0.117111	0.045137	1.000000	0.027261	0.171999	0.100411	0.000000	0.000000	
TEERNPER_5.0	0.191691	0.075984	0.027518	0.027261	1.000000	0.106706	0.070491	0.005136	0.000000	
TEHRUSLT_25.0	0.576347	0.427135	0.163306	0.171999	0.106706	1.000000	0.487810	0.210142	0.175824	
TEHRUSLT_40.0	0.325572	0.254514	0.065357	0.100411	0.070491	0.487810	1.000000	0.130236	0.108789	
TEHRUSLT_50.0	0.189185	0.182306	0.042452	0.000000	0.005136	0.210142	0.130236	1.000000	0.045098	
TEHRUSLT_45.0	0.149678	0.149754	0.027255	0.000000	0.000000	0.175824	0.108789	0.045098	1.000000	
TEHRUSLT_60.0	0.092265	0.079504	0.029840	0.000000	0.017777	0.136091	0.083897	0.033780	0.027181	
TEIO1COW_4.0	0.167650	0.150416	0.043461	0.048876	0.101198	0.369523	0.159014	0.063003	0.021772	

We used One Hot Encoding to help with our largely categorical data set. The purpose of One Hot Encoding is to map various categorical variables like the ones in our data to integers that can be better used and understood by machine learning algorithms. This is particularly useful when working with data that may not all be related. It also helps when working with integers because of the way machine learning treats number order (Higher is better), as well as allowing for better scalability.

Zero values were previously addressed using mean value imputation. Some of our data contained 0's or other indicated non values. We chose to use Mean Value Imputation to account for these zeroes. Mean Value Imputation does exactly what it sounds like, it replaces the non-values in our data, for example 0, with the mean value of all the values in the column the data is found in.

Methodology

Throughout this project, we adhered to the Cross Industry Standard Process for Data Mining, or CRISP-DM methodology as a guide for our processing and decision making. Keeping CRISP-DM in mind, we focused primarily on using the Python programming language, specifically the Pandas library to work with and shape our data into a form that could be used to make meaningful determinations about the observations contained in the data.

We opted for an unsupervised learning approach to our data, specifically clustering. We chose this method because with unsupervised learning, you can find relationships in data that aren't immediately apparent. It does this by comparing the data based on similarity. In clustering, this is used to group the unlabeled data into clusters that represent patterns that were found in the data.

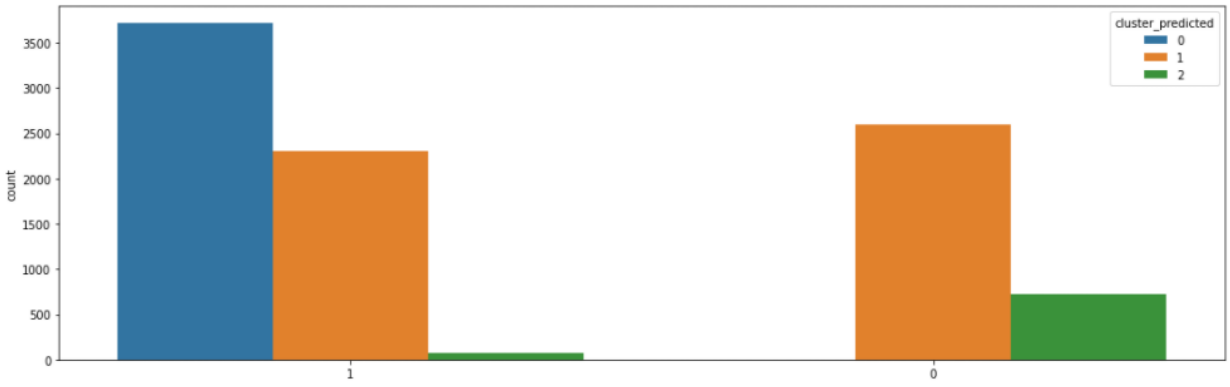
Our focus was using K Mode clustering, an extension of K-Means clustering, to determine the relationships that exist between our variables by presenting them as a group of visually clustered data points which reveal relationships not immediately apparent from the data set. The number of clusters is determined by the K value provided to the method, with several ways to determine the correct K value including an

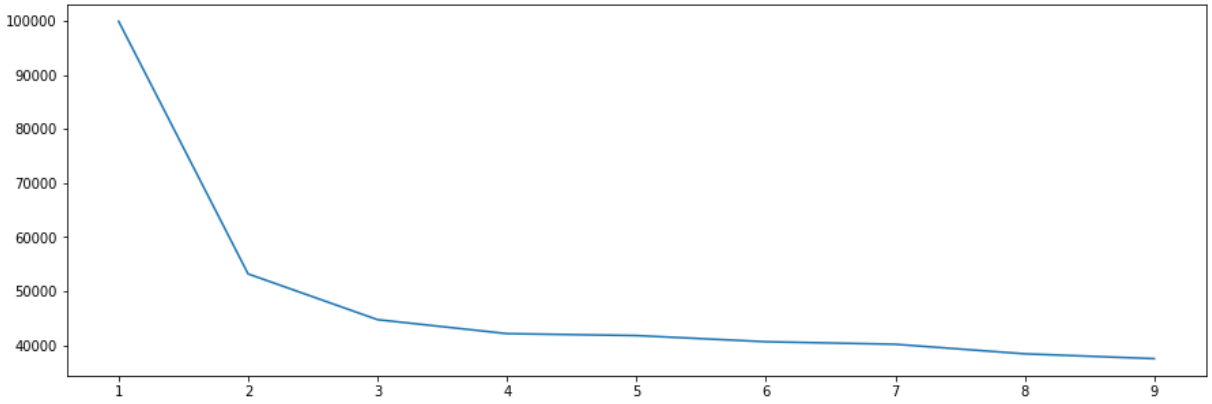
Elbow Chart or a Silhouette Graphs that can provide a visual representation of ideal K values for our data set.

We chose K-Mode clustering because of the nature of our data. K-Mode clusters are extremely efficient when working with large amounts of categorical data. K-Mode doesn't need to calculate all of the pair wise distances between data points. Additionally, it is distribution free, meaning that it does not require imposing distribution assumptions on the data in order to work.

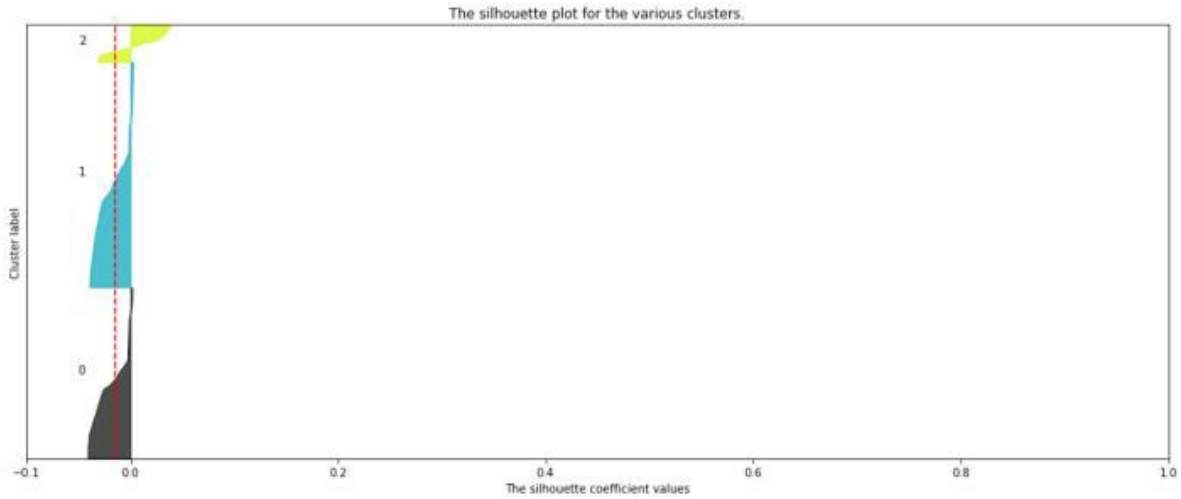
K-Means clustering on the other hand does not work well with categorical data sets, primarily because it has the issues above that K-Mode does not, and ultimately, its objective function simply doesn't work with categorical data.

Clustering is a means of grouping data based on characteristics. The clusters will form based on relationships between the individual data points and potentially reveal new relationships that were not previously apparent. Clustering also allows for data compression when working with large data sets. After the data is clustered it can be referred to by cluster ID rather than the individual data points themselves. Another use of clustering is the preserve the privacy of the respondent. Data grouped in clusters is practically anonymized by its presence in the cluster and future reference by cluster ID. In our example, no individual respondent ID will need to be used going forward as we can refer to the groups by their cluster IDs.


cluster_predicted
0
1
2
count
3500
3000
2500
2000
1500
1000
500
0
1
0


100000
90000
80000
70000
60000
50000
40000
1
2
3
4
5
6
7
8
9

bends, but the one with the least dispersion associated with it is at K=3, and as such that was the value we went forward with.



The silhouette Method is also a method to find the optimal number of clusters and interpretation and validation of consistency within clusters of data. The silhouette method computes silhouette coefficients of each point that measure how much a point is similar to its own cluster compared to other clusters. by providing a succinct graphical representation of how well each object has been classified.

Silhouette analysis can be used to study the separation distance between clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters.

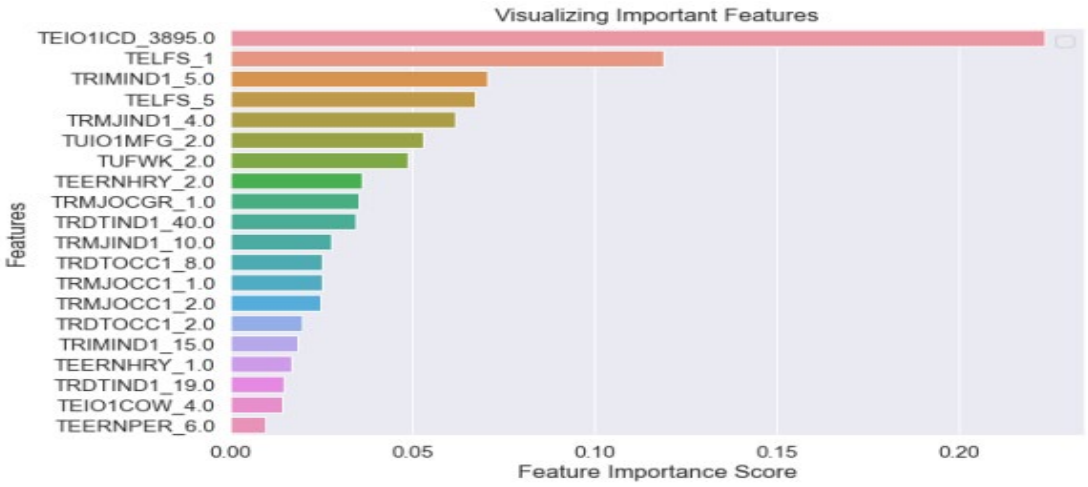
Observations from above Silhouette Plots

- The silhouette plot shows that the n cluster value of 3 is a bad pick, as all the points in the cluster with cluster label = 0 are below-average silhouette scores.
- The silhouette plot shows that the n cluster value of 5 is a bad pick, as all the points in the cluster with cluster label = 2 and 4 are below-average silhouette scores.
- The silhouette plot shows that the n cluster value of 6 is a bad pick, as all the points in the cluster with cluster label = 1,2,4 and 5 are below-average silhouette scores, and also due to the presence of outliers.
- Silhouette analysis is more ambivalent in deciding between 2 and 4.
- The thickness of the silhouette plot for the cluster with cluster label = 1 when n clusters=2, is bigger in size owing to the grouping of the 3 sub-clusters into one big cluster.
- For n clusters = 4, all the plots are more or less of similar thickness and hence are of similar sizes, as can be considered as the best value for ‘k’.

Findings

- Based on Employment Survey response from individuals, it seems our unsupervised Learning model suggested to divide the individuals
  in three different groups (clusters).
- First group of individuals are seeming to have full employment in manufacturing sectors
  (most of them are serving Furniture and fixtures manufacturing).

- Most of the individuals belongs to first group are having Management, professional, and related occupations In the similar way our model had decoded the characters for other groups too. We have used unsupervised learning to predict the clusters and its accuracy is very low, so we would need to use additional Ensemble techniques to increase its accuracy



Business Application

The goal of this project is to assist in the marketing of financial services to customers based on self-identified demographic data from the Bureau of Labor Statistics, allowing for the creation of pools of demographically sorted candidates for client services.

Based on our findings, we can craft tailored survey data to better target ideal customers for our clients. This goal is met by creating a predictive model to determine if a given group of respondents is likely to be a good fit for our clients' financial products and as a result an ideal target for marketing campaigns attempting to solicit customers for those products.

Acknowledgements

We would like to thank our fellow students and classmates for being there for us to bounce ideas off of. We would also like to thank our professor, Dr. Fadi Alsaleem for his mentorship and guidance throughout the process of working on this project. Finally, we would like to acknowledge the Bureau of Labor Statistics for collecting the data we chose to work with as well as the individual respondents for providing answers to the questions asked in the survey.

References

- Data Science in 5 minutes: What is one hot encoding? Educative. (n.d.). Retrieved February 9, 2022, from https://www.educative.io/blog/one-hot-encoding

- Decision trees in python - step-by-step implementation. AskPython. (2020, December 7). Retrieved February 10, 2022, from https://www.askpython.com/python/examples/decision-trees

- Franklin, S. J. (2019, November 26). Elbow method of k-means clustering algorithm. Medium. Retrieved February 9, 2022, from https://medium.com/analytics-vidhya/elbow-method-of-k-means-clustering-algorithm-a0c916adc540

- Google. (n.d.). Clustering algorithms | clustering in machine learning | google developers. Google. Retrieved February 10, 2022, from https://developers.google.com/machine-learning/clustering/clustering-algorithms

- Jaiswal, S. (2020, July 10). K-modesclustering. Medium. Retrieved February 10, 2022, from https://medium.com/@shailja.nitp2013/k-modesclustering-ef6d9ef06449

- Mean imputation for missing data (example in R & SPSS). Statistics Globe. (2022, January 19). Retrieved February 9, 2022, from https://statisticsglobe.com/mean-imputation-for-missing-data/

- Selecting the number of clusters with silhouette analysis on kmeans clustering. scikit. (n.d.). Retrieved February 9, 2022, from https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

- U.S. Bureau of Labor Statistics. (n.d.). Atus News releases. U.S. Bureau of Labor Statistics. Retrieved February 10, 2022, from https://www.bls.gov/tus/