**Predicting number of hours worked and other classification by Gavin Whitesitt**

**1. Problem Statement, Motivation, Research Goals:**

The number of hours in a week that we work determines much of our life experience. It largely determines how much time we spend away from our family. It determines what we think about for a large portion of our life. It can also determine if we are successful or not. Given a set of characteristics of a person it would be interesting to be able to predict how long they work. Finding characteristics that are associated with long and short work hours might be interesting in better understanding why people put in long hours and why people don’t put in long hours. The goal of this research is to predict the number of hours that someone works based on a set of their characteristics.

**2. Data Source and Description:**

Where/how do you get the data?

I accessed the data for usa.ipums.org I used their select data tool to construct the dataset used in analysis. The sample I used is from 2016.

**3. Literature Review and References:**

This could include noting any key papers, texts, or websites that you have used to develop your modeling approach, as well as what others have done on this problem in the past. You must properly credit sources.

Python Data Science Handbook by Jake VanderPlas

Introduction to Machine Learning with Python by Andreas Mueller and Sarah Guido

“Fast algorithms for mining association rules” by Agrawal, Rakesh, and Tamakrishnam Srikant

“[Clustering] How to sort a distance matrix” by gautier marti

Finding Groups in Data: An Introduction to Cluster Analysis. By Kaufman, Leonard, and Peter Rouseeuw

An Introduction to Statistical Learning with Applications in R by Gareth james Daniela Witten Trevor Hastie and Robert Tibshirani

**4. Preliminary EDA:**

**a) What is the shape of your data set - how many rows (observations), and** how many columns (variables/features)?

The dataset has 3,156,487 rows and 259 columns.

**b) What are the names of these variables, and its full descriptions?**

The name and full description of the variables can be found [Here](https://usa.ipums.org/usa-action/downloads/extract_files/usa_00007.xml%23HHWT). It can also be found in an xml file on my github.

**c) How many numerical variables?**

The dataset has 20 continuous numerical variables

**d) How many categorical variables?**

There are 239 categorical variables

**e) How many text variables?**

There are no text variables

**f) What variables other than the above are involved?**

There are only numerical and categorical variables involved

**g) What methods have you used to preprocess/clean/explore the data and why?**

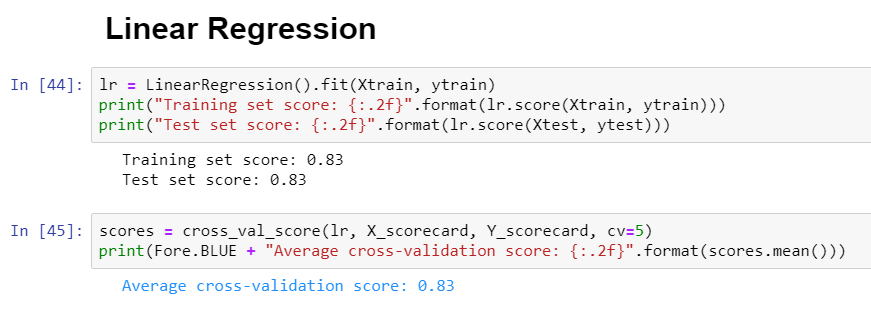
The data is already in a pretty nice format. I dummy coded all the categorical. I changed all values specified in the documentation as NA’s to actual NA’s. After this I dropped any rows with any NA’s. Due to limitations of my hardware I took a 1% random sample from the large dataset and performed Linear Regression, K Nearest Neighbors Regression, Lasso Regression, Ridge Regression, Decision Tree Regression, Ada Boost Regression, Extra Trees Regression, Random Forest Regression, and Bagging Regression. Out of these the initial findings suggest that the Bagging Regressor is performing the best explaining 88% of the variance.

You will provide related visualizations, summary statistics, and verbal descriptions.

**5. Modeling Process (the main portion):**

Provide a reproducible modeling process (with all codes and comments, from Data to Models) of fitting

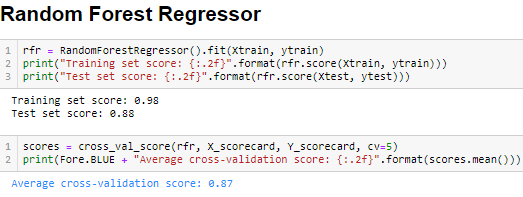
All code for reproducibility is attached in later pages of this document.

a) an initial baseline (simple) model for comparison (regression)

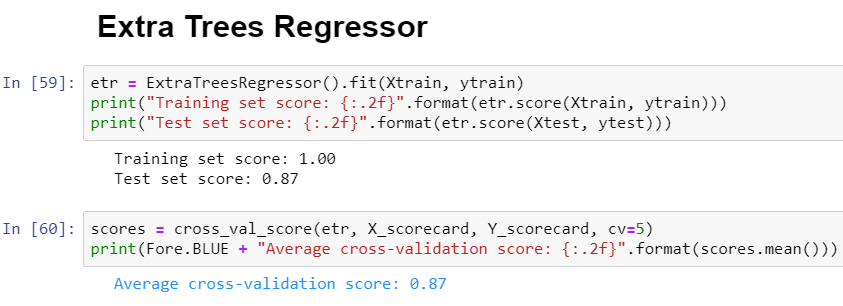
b) a set of competitive feasible models (regression)



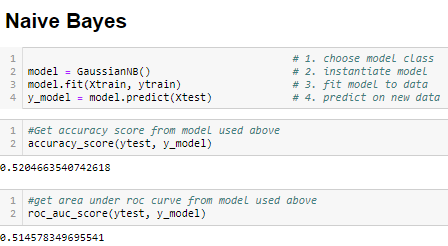




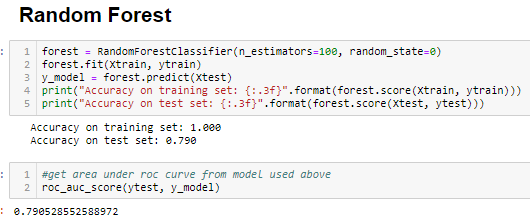
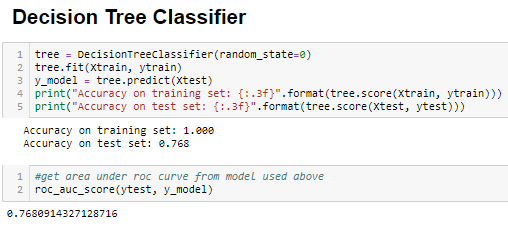
c) a final model of your best choice (regression)

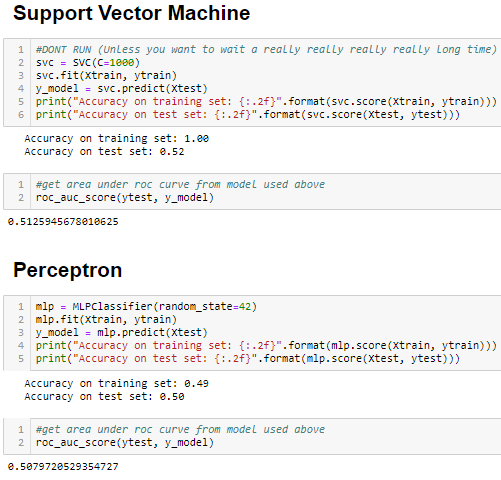


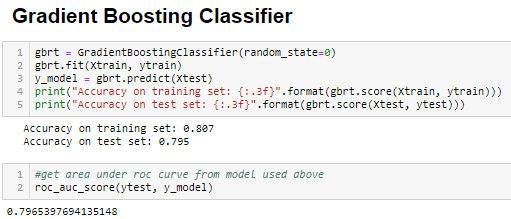
1. an initial baseline (simple) model for comparison (classification)



1. a set of competitive feasible models (classification) 





1. a final model of your best choice (classification) 

**6. Project Progress, Timeline, and Achievement:**

Briefly report your progress with a clear timeline (in terms of the number of weeks of the semester) and summarize your project achievement step-by-step that you have made along the way from beginning to end.

This timeline will tell how you arrived at your results and powerfully illustrate your efforts in the due process. How well does your model and/or implementation perform? Did you meet your goals? What are the significance of your results?

**Week Of 10/21**: Chose data set. Developed Research question. Performed Preprocessing: Imputed missing values, encoded categorical variables, dropped some redundant features. Split data into test train set. Performed initial analysis with 1% sample. Created baseline linear regression model for future comparison.

**Week Of 10/28**: created new models for comparison of the original baseline (Ridge Regression, K Nearest Neighbors Regressor, Lasso, Decision Tree Regressor, Ada Boost Regressor, Extra Trees Regressor, Random Forest Regressor, Bagging Regressor). Added five-fold cross validation. Began writing code for unsupervised learning section (Clustering, Seriation).

**Week Of 11/4**: Implemented unsupervised learning methods (Principle Component Analysis, Determined Optimal Number of Clusters using BIC vs AIC curves, Elbow Method)

**Week Of 11/11**: Further implemented unsupervised learning methods (Spectral Clustering, Gaussian Mixture Modeling, Agglomerative Clustering, DBSCAN, MiniBatchKmeans). Added Seriation.

**Week Of 11/18**: Started to implement association rules. Encoded certain values that should have originally been coded as NA.

**Week Of 11/25**: Scaled dataset before PCA. Adjusted analysis accordingly

**Week Of 11/31**: Added classification methods (Naïve Bayes, Logistic Regression, K nearest neighbors, decision tree classifier, random forest, gradient boosting classifier, support vector machine and peceptron) based off of Sex as the target variable.

**Week Of 12/3**: Discretized continuous variables, added more association rules. Watched my computer crash several times due to memory errors while creating association rules :) . Created final project powerpoint.

**Week Of 12/10**: Re-ran all regression and classification with scaled dataset. Finalized Final Project

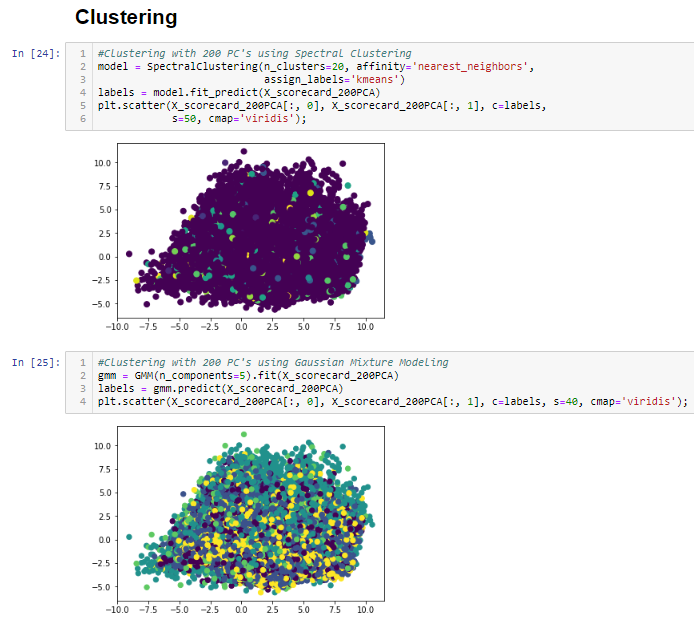
**7. Conclusions and Possible Future Work:**

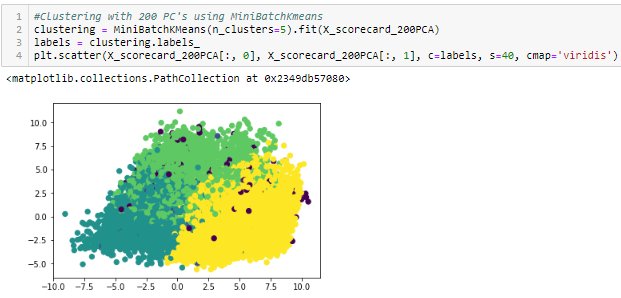
**Summarize the strengths and weaknesses of your results, speculate on how you might address these short-comings, and plan for further research directions if given time:**

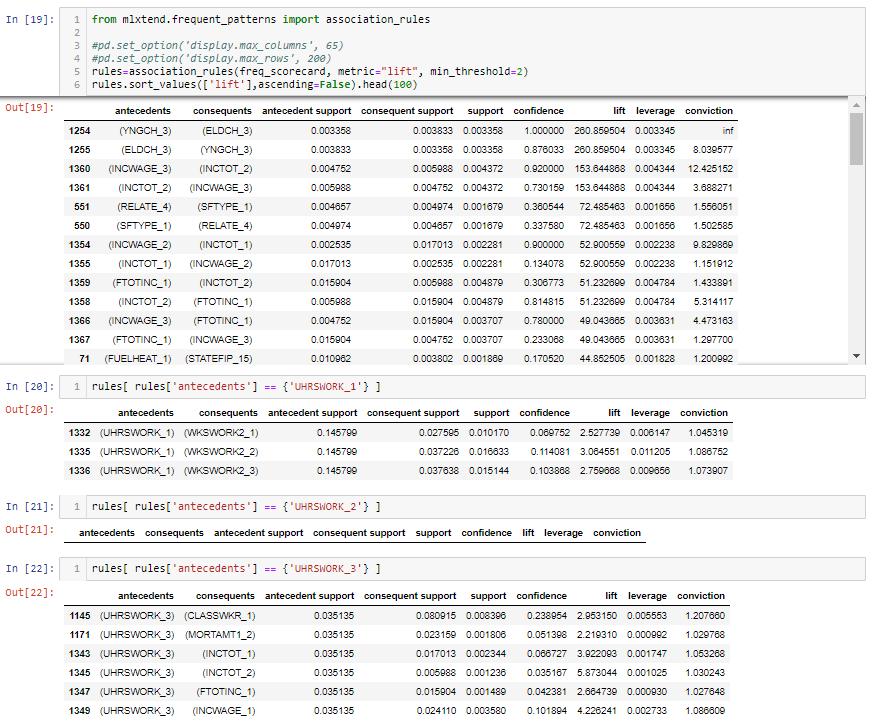
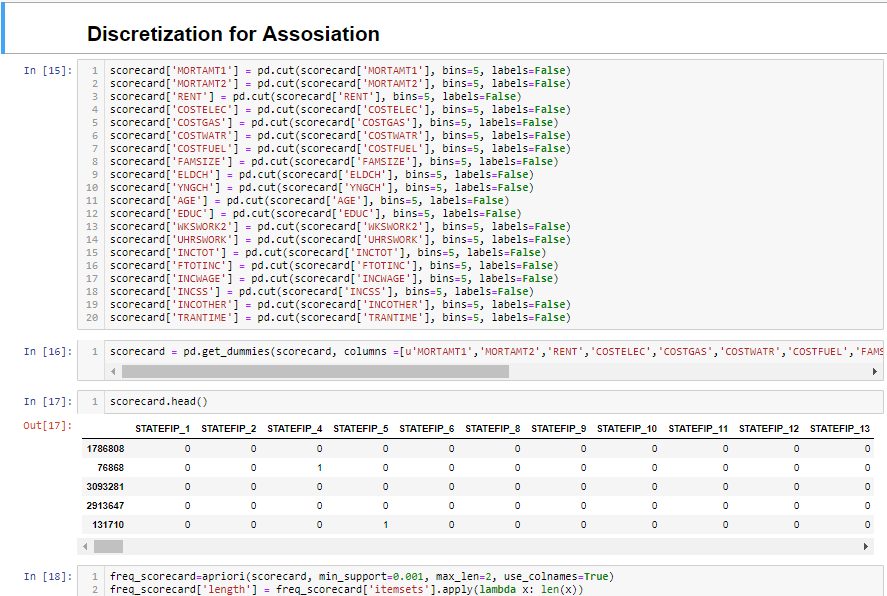
The results are less positive than I originally hoped to accomplish with this project, but I hope to continue to work at the problem in the future at some point to answer my original research question better. A relatively large percentage of the variance can be explained in only a few principle components. The weakness of my results lie in the fact that only 1% of the data available for 2016 is used, it is possible that if more computational power is used my model could be improved. In the future I would like to use data from multiple years to determine how association rules have changed over time as well as how a time series dataset would affect the prediction accuracy of my models.

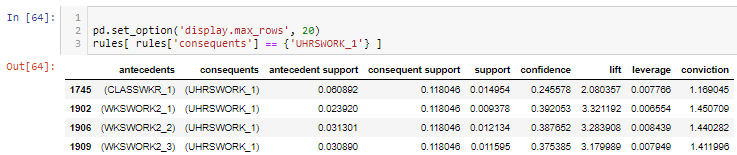
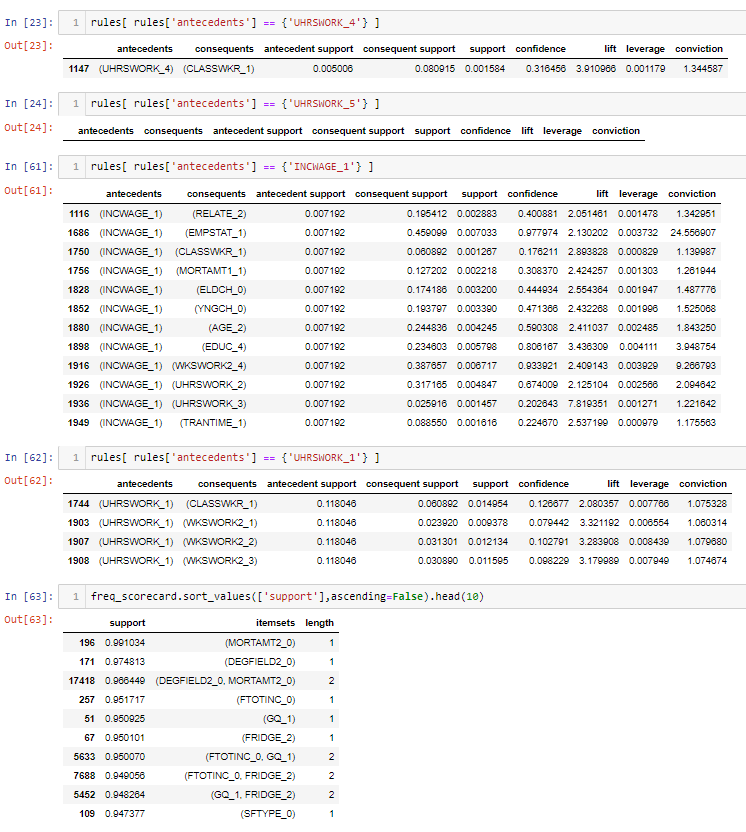
Overall, I think there is a lot of room for improvement. The main bottleneck was my computational hardware. In the future as computational power increases and becomes more affordable I would like to re run my analysis with a larger subset of the data. This should also allow me to mine for more interesting assosiation rules. I would say that I did ultimately meet my goals in regards to the implementation of the code that would allow me to answer the questions posed. It just happened to be a little too compuationally difficult to run the code on my hardware. I feel as if this project Is just the first in a series of projects attempting to answer similar questions.

**Additional Output:**

****

****





\*A set of the entire reproducible code can be found at my github at github.come/Gavin-Whitesitt