PREDICTING INCOME LEVELS

Francesca Schott
Course Project

General Assembly
Part Time Data Science
Fall 2016

CONTENTS

1 INTRODUCTION

American's spend their time in distinct and wide-ranging ways over the course of a day. However, as is often the case, for all the ways in which we are unique, it can be shown that we are not as different as we think.

The structure of a person's 24-hour day provides insights into their occupation, household makeup, eating habits, health and wellbeing, and many other factors that, in total, are the formation of our daily lives. When time use is examined closely, there is an opportunity to evaluate and make predictions based on similarities. While evaluating the connections between American time use and income levels could be conducted from either direction, this project explores time use variables, informed by survey data collected by the United States Bureau of Labor Statistics, to predict American's income levels.

Ultimately, if certain time uses and lifestyle choices are found to positively or negatively affect income levels, an outcomes-based argument for raising income levels may encourage people to make better choices about their health, eating habits and overall wellbeing.

2 STATEMENT FOR ANALYSIS

Evaluate and try to predict the income level of Americans, as categorized by the United States Census Bureau, using information about exercise, food availability, grocery shopping locations and job type.

3 DATA

The United States Bureau of Labor Statistics (BLS) performs an annual survey, known as the American Time Use Survey (ATUS), to collect data about how Americans spend their time. The survey covers a broad range of topics from household size to eating habits to time spent caring for elderly relatives.

As an addition to the ATUS, the BLS partners with the United States Department of Agriculture's Economic Research Service to conduct additional survey questions of a select group of ATUS respondents about their eating and health habits. The Eating and Health Module (EHM) of the ATUS is not conducted annually and the most recent data files available are from 2014.

Using data from both the 2014 ATUS and 2014 EHM, I wanted to explore how respondent's income might be predicted using various inputs.

3.1 Merging & Creating Subset

All the data sets associated with the ATUS contain linking variables specifically provided to encourage users to link data sets based on the project or questions being asked. Using the linking variables, I merged the following columns from the ATUS Respondent Files onto the EHM Respondent Files:

- Weekly income
- Major industry group
- Occupational group
- Labor status
- Household number

The weekly income feature was chosen as a possible prediction variable, the major industry group and occupational group were chosen to use as possible features for predicting income levels, and labor status and household number were chosen to reduce the total number of responses provided in the data set to analyze a subset of only employed, single-person households.

In addition to employed, single-person households, respondents answered questions about their income based on either 2014 or 2015 poverty threshold levels. The yearly poverty threshold variable was tracked by the survey collectors and included in the EHM data set. I chose to analyze only responses based on the 2014 poverty thresholds. Of the 1,158 responses left after reducing the data set, eight responses about income level were unanswered. I dropped the 8 null responses for a total response count of 1,150, the final response count for the data set used for predictions.

Once the final response count was set, I cleaned and organized the final column names, ultimately choosing nine variables to explore for feature selection. More information about the final data set can be found in the data dictionary here. The final columns names were:

Column Name	*Brief Description*	*Variable Type*
bmi	Body Mass Index	Continuous
primary_eating	Time spent eating in primary	Continuous
secondary_eating	Time spent eating while engage in other activities	Continuous
exercise	Exercise performed in past 5 days	Categorical
fast_food	Fast food purchased in the past 7 days	Categorical
food_amount	Enough food to eat in past 30 days	Categorical
stores	Where majority of groceries are purchased	Categorical
maj_ind	Major industry based on main job	Categorical
cat_occ	Occupation category based on main job	Categorical
income_lvl	Income level based on poverty thresholds	Categorical
income_weekly	Weekly income based on hourly wage and hours worked per week	Categorical

With the nine final features selected, I worked to clean these columns by replacing null values with zero for categorical valuables and the mean value for continuous variables. There were also three types of negative-value responses tracked by the survey to categorize types of refusals to respond. Disregarding the type of refusal, I chose to make any negative categorical responses zero for categorical variables and equal to the mean for continuous variables.

Finally, for the response variables, I could choose to predict either the continuous variable of weekly earnings or the categorical variable of income level based on poverty thresholds. I chose to use the categorical variable of income level as my predictor. The EHM column for income level was given in five categories based on two levels of the poverty threshold, 185% and 130%. I decided to create a binary variable for predictions and assigned two income level responses with a value of zero or one, either below 185% of the poverty threshold level or above 185% of the poverty threshold level.

3.2 Feature Selection & Engineering

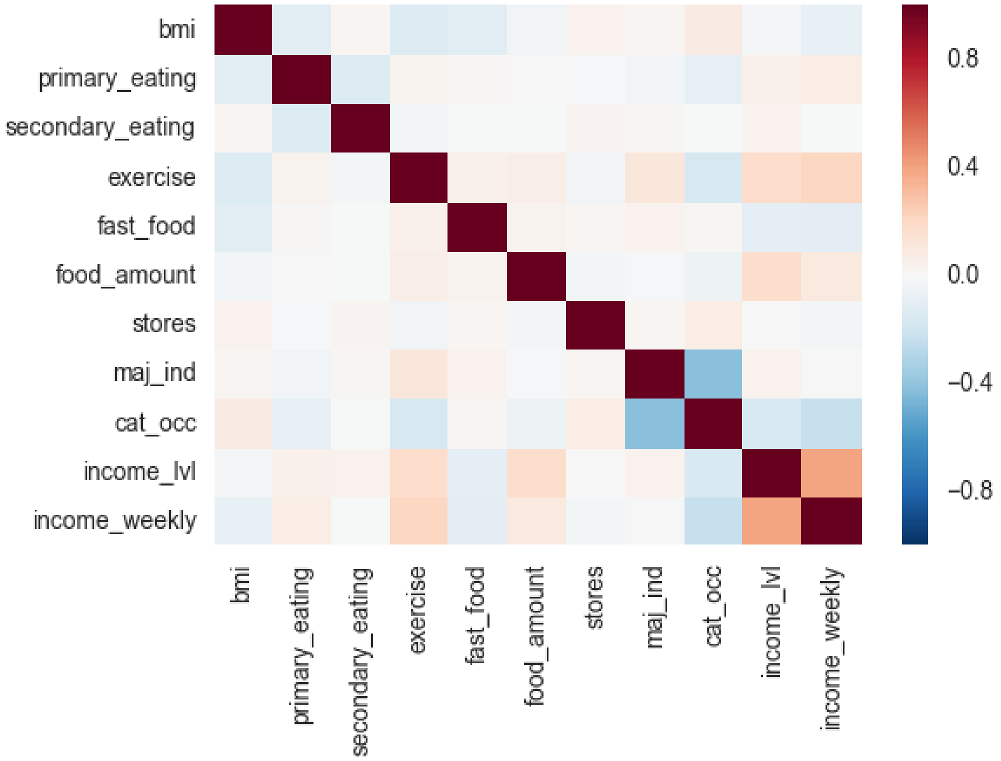
The original EHM data set started with 43 columns. Some of these columns were related to the linking variables used across all ATUS data sets and were easily dropped. For feature selection, I initially used the EHM data dictionary to review the questions asked and way the response was recorded. I chose to focus on about 20 features that were informative in how the question was asked and were generally complete without many missing values.

With about 20 features selected, after several iterations of exploration using various data description methods, correlation heat mapping, and plots for visualization, I narrowed the set down to the nine columns described in section 3.1 above.

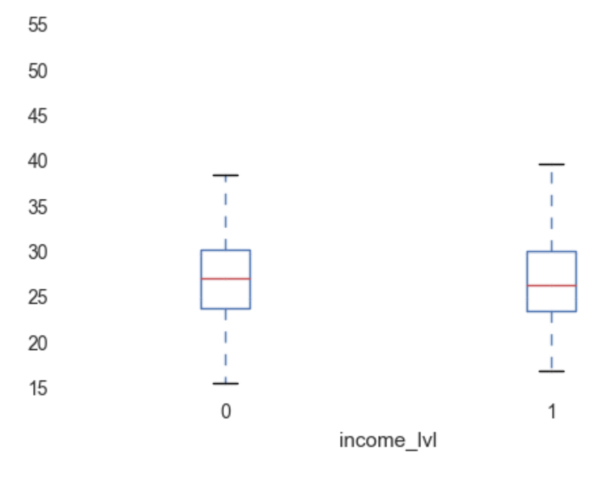
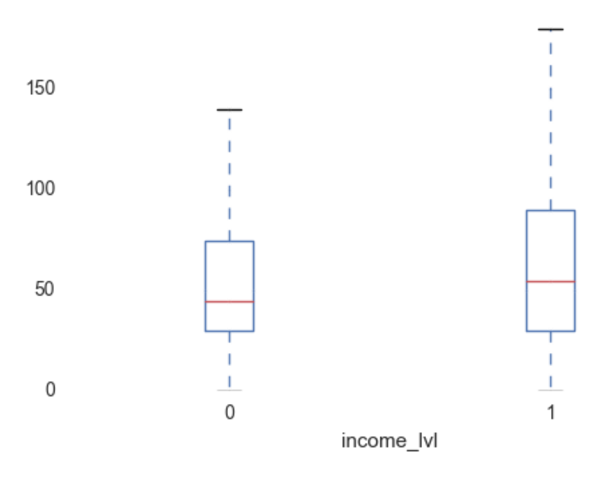
3.3 Exploration & Visualization

Note: Please reference the data dictionary for complete description of the features graphed in this section.

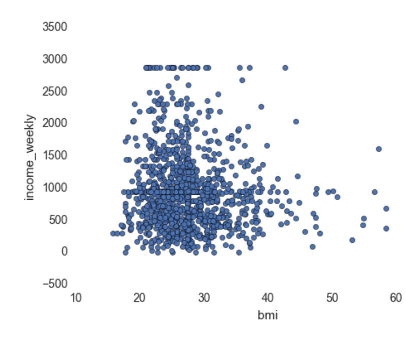
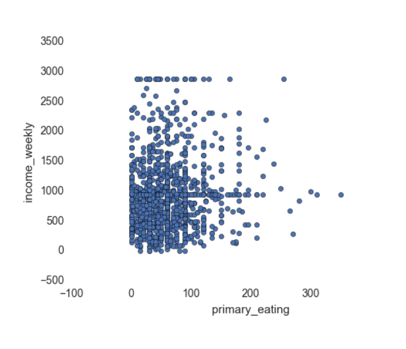
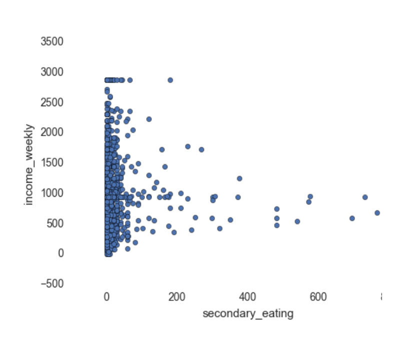
After working to clean and engineer my data, I created a correlation heat map to use as a starting point and as a benchmark to visually explore my features. Shown below is the heat map with the nine selected features and the two predictors.

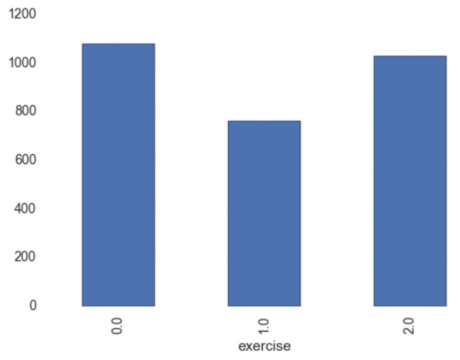
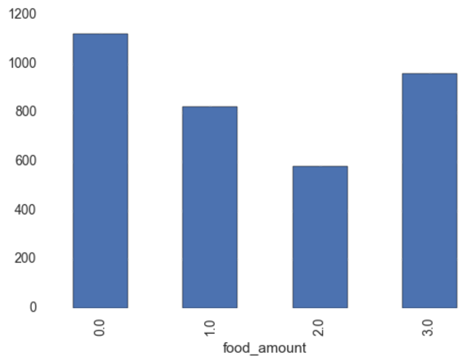
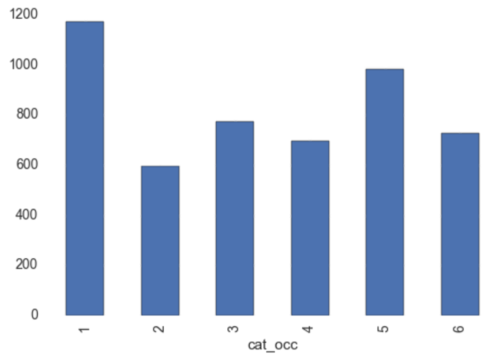
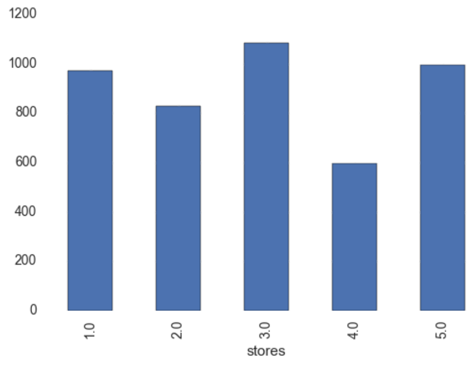


As seen below in the box plots using income level for grouping, the body mass index and primary eating variables do not vary significantly for someone making below (0) or above (1) the 185% poverty threshold income level. After reviewing the correlation heat map, in addition to models functioning poorly when including these two variables, they were not selected as features to use in predicting income levels.

Box Plot Grouped By Income Level Describing BMI	*Box Plot* Grouped By Income Level Describing Primary Eating Time
	

Because income level and weekly income are strongly correlated variables, I chose to use weekly income as a continuous variable to visualize several features from the data set. Below are the scatter plots and bar charts visualizing body mass index, primary eating time, secondary eating time, exercise, food amount, occupational category and stores.

Scatter Plot BMI x Weekly Income (in dollars)	*Scatter Plot* Primary Eating Time x Weekly Income (in dollars)	*Scatter Plot* Secondary Eating Time x Weekly Income (in dollars)
		

Bar Chart Exercise x Weekly Income (in dollars)	*Bar Chart* Food Amount x Weekly Income (in dollars)
	
Bar Chart Occupation Category x Weekly Income (in dollars)	*Bar Chart* Stores x Weekly Income (in dollars)
	

4 MODELING

As a classification problem trying to predict someone's income above or below the 185% poverty threshold, several modeling options were available. All the models used were attempting to improve the null accuracy value of .755652.

For simplicity and due to time constraints (simply my lack of planning combined with life-imposed timing issues), all my models were basic models, run without in-depth feature selection or parameter tuning. As I continue to work on improving this project and its results, addressing feature selection and parameter tuning will be my next step.

I created and used the same test-train-split for all the models, with scaled OneHotEncoder dummy variables used in the K-Nearest Neighbors (KNN) and logistic regression models, and measured the model accuracy using a 5-fold cross validation.

In addition to the classification modeling used to predict income levels, I performed some initial clustering analysis on the weekly income continuous variable. The results of each analysis are summarized in section 4.1 and 4.2.

4.1 Income Level Classification

<table>
<tr><th><i>Model</i></th><th><i>5-Fold Cross Validation Accuracy Score</i></th><th><i>Additional Analysis/Notes</i></th></tr>
<tr><td>KNN</td><td>0.79224</td><td></td></tr>
<tr><td>Logistic Regression (LogReg)</td><td>0.77783</td><td>Confusion Matrix:<table><tr><td>14</td><td>55</td></tr><tr><td>8</td><td>211</td></tr></table></td></tr>
<tr><td>LogReg Pipeline</td><td>0.80259</td><td>SelectKBest feature selection</td></tr>
<tr><td>Decision Tree</td><td>0.78534</td><td><table><tr><th></th><th>Feature</th><th>Importance</th></tr><tr><td><b>0</b></td><td>exercise</td><td>0.219443</td></tr><tr><td><b>1</b></td><td>food_amount</td><td>0.027940</td></tr><tr><td><b>2</b></td><td>cat_occ</td><td>0.736690</td></tr><tr><td><b>3</b></td><td>stores</td><td>0.015926</td></tr></table></td></tr>
<tr><td>Bagged Decision Tree</td><td>0.75369</td><td></td></tr>
<tr><td>Random Forest</td><td>0.75714</td><td></td></tr>
<tr><td>Voting Classifier</td><td>0.78472</td><td>LogReg, Decision Tree, Random Forest</td></tr>
</table>

4.2 Weekly Income Clustering

As an initial investigation into the weekly income variable, I chose to perform a k-means cluster analysis using five clusters. The Silhouette Coefficient was calculated to be 0.57528, a relatively strong value considering the simple feature selection methods that were used.

In the table below, I have highlighted the first and fourth features as interesting results to review. The most noticeable difference, compared to the other three clusters, is the exercise feature. The first and fourth cluster were the only two that strongly categorized people who did exercise (2) and people who did not exercise (1) in the past five days of being asked by the survey.

With this drastic difference from the other three clusters, we then note that the occupation category is quite different from the other three clusters. By reviewing the bar chart in section 3.3 we see that occupation category ‘1’ and occupation category ‘2’ vary the most of any occupational categories. While not as apparent a difference as the exercise feature, combined with the weekly income amount and the strong divide in the exercise feature, an argument could be made that the occupation category in the first and fourth clusters does contribute to how the cluster was formed.

Cluster Number	*Exercise*	*Food Amount*	*Stores*	*Occupation Category*	*Weekly Income (in dollars)*
0	2.000000	3.000000	1.228477	1.587748	1067.923626
1	1.572254	3.000000	1.358382	5.595376	826.925137
2	1.641509	2.981132	4.905660	2.566038	971.883505
3	0.992157	3.000000	1.258824	1.890196	782.221912
4	1.523077	1.646154	1.461538	2.707692	654.718915

5 CHALLENGES

Understanding which models performed well, and which features contributed to overall model performance, took the most time and required the most extensive review of the materials provided throughout the 10 weeks of classes.

Additionally, because the in-class examples used clean data sets with strong predictors, such as the iris data set, it was challenging to decide if a correct, strong improvement in predictions was achieved with each feature cleaning and selection process. Using clean, strongly-predicting data sets is understandable for the examples in class, and provided a strong platform to begin using real world data, cleaning it, and then deciding if the accuracy at which it predicted the outcome was acceptable. Our tenth class, the midpoint of the course, was a review day, and was incredibly helpful both as a general review and in working with a never-before-seen data set.

6 NEXT STEPS

With so many possibilities provided by the breadth of the ATUS data set, I am interested in pursuing several next steps to refine the data subset used in this project and to expand the set to incorporate other features and perhaps other predictors.

Initially, for the current subset of employed, single-person households responding within the 2014 poverty threshold, I want to complete more thorough feature selection and parameter tuning for individual models. I did not implement a grid search for any of my models and I will start there to being

Also with the current subset, I have developed a rough Heroku App that I will continue to clean and format as I improve my modeling. The app can be viewed here: www.income-pred.herokuapp.com With improved funcationality, this app could be used as a quick tool for anyone in the health and/or wellness fields to demonstrate the lifestyle differences that contribute to differences in income levels.

Beyond the current subset of data, I want to investigate if there are stronger features that better predict income levels, broaden my subset to include multi-person households, and explore if there other features and/or predictors that ultimately create strong arguments and could be used to encourage people to lead healthy, well-balanced lives.

7 CONCLUSION

The complexity and exhaustiveness of the ATUS data set presents an opportunity to explore many other questions associated with how Americans spend their time. Reflecting back on my process through this project, I am not sure that my overall decision to try and predict income levels was correct for the data set I chose. In the future, I will work to understand this data set more thoroughly and try to take advantage of its annual release.

Throughout the data cleaning and exploration I realized that features I thought would be strong predictors were not, and unexpected features began to emerge as more useful. As part of the learning process, as much as I seem to forget, any preconceived notion of how to structure a project (or, perhaps, a person’s daily schedule) often must adjust to unexpected influences, outside factors, and the overall unpredictability of life itself. So with the subtle irony of trying to predict outcomes based on the unpredictable factors of how we spend our days, I look forward to pulling from these lessons learned as I continue to grow as a data scientist and as a person.