Predicting Adult Income with Adult Data Set CSI 5810

Adam DeJans

November 11, 2017

Date Presented: November 7, 2017
Instructor: Professor Sethi

The following report is arranged as five parts. The first section gives the objective and some background information on the problem at hand. In the second section, the dataset is introduced. Some important features of the dataset are discussed along with background knowledge and some interesting visualizations of the dataset are shown. The features extracted from the dataset that are used in the tested models are also discussed in the second section as well. In the third section, classification algorithms used/tested are discussed. The performance amongst different parameters are also seen in the third section. In the fourth section, the experiment (testing the model on the provided test dataset) setup and result is introduced. In the fifth section, the conclusion is given and some possible ways to improve the model is concerned.

1 Objective and Background Information

The objective of this project is to determine whether an adult person makes more or less than fifty-thousand United States Dollars per year using some given attributes of an individual. The data was extracted from the 1994 Census bureau database. For each row in the data there are fourteen attributes, both continuous and categorical, that are measured.

It has been found through many sociological texts that when it comes to wages there is a race and gender gap. It has been studied that given similar occupations: Asians make the highest income (with whites as the second clear highest), and men make more than women. We have also researched that the middle-aged people, defined as ages $35 - 54$, also make a higher average income as compared to their younger and older counterparts. This makes sense as young people are

just starting their careers and elderly people are beginning to retire.

When proceeding with the project we will consider this background information, especially when we reach the point of feature extraction.

2 Dataset Analysis

The dataset was donated by Ronny Kohavi and Barry Becker of Silicon Graphics. The data was extracted from the 1994 Census database.

The dataset provided was split into two files: the training set *adult.data* and the testing set *adult.test*. We only explore in the adult.data, after which we reduce/clean it and do a k-fold cross validation (later discussed in the fourth section). The adult.data dataset originally had 32,562 rows of data. Each row in the dataset represented a description of an single adult from the 1994 Census database.

Here is a sample of what the data looked like originally:

	Age	Workclass	fnlwgt	Education	Education-Num	Martial Status	Occupation	Relationship	Race	Sex	Capital Gain	Capital Loss	Hours per week	Country	Target
30156	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	United-States	<=50K
30157	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-States	>50K
30158	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	United-States	<=50K
30159	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-States	<=50K
30160	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	United-States	>50K

Figure 1: A subset of the dataset.

Within the dataset we see that there is a mixture of continuous and categorical data:

Continuous Variables	Categorical Variables
Age	Race
Capital-gain	Sex
Capital-loss	Workclass
Fnlwgt	Native-country
Hours-per-week	Relationship
Education-num	Education
	Marital-status
	Occupation

We first plot the distributions of all variables and the classes:

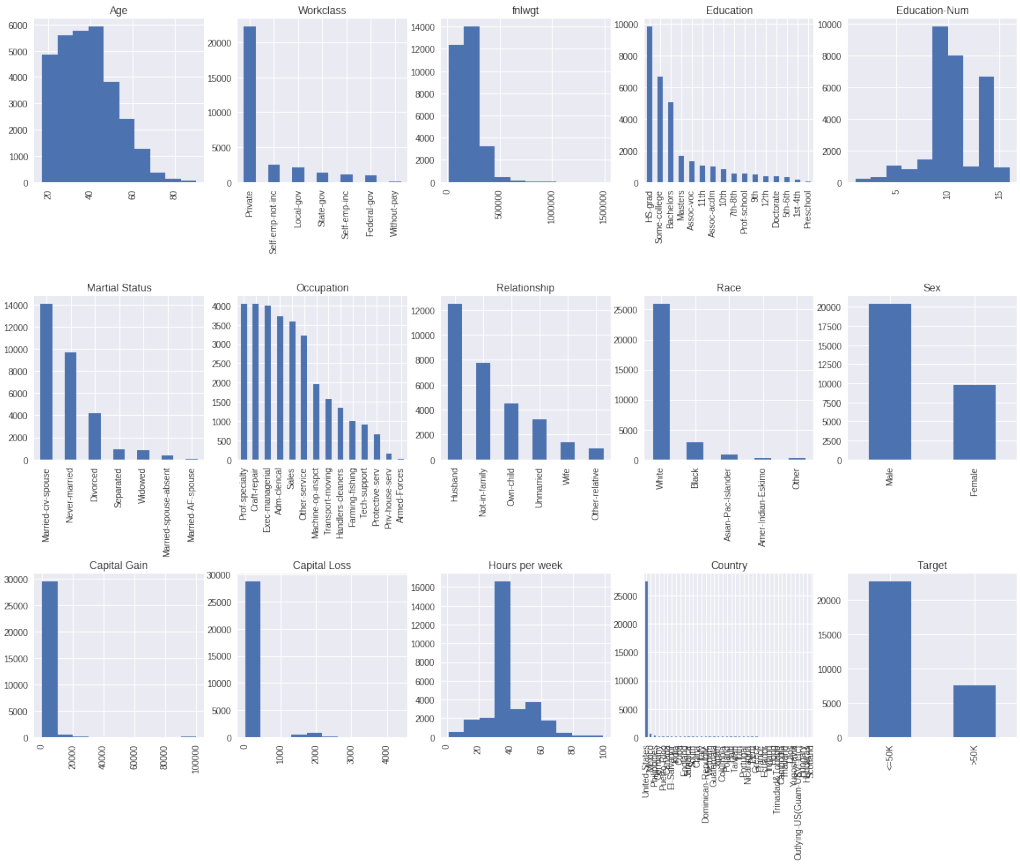

Figure 2: Distribution across all variables.

The first thing to notice is the class distribution of the data. The class distribution is strongly biased towards a person making less than or equal to fifty-thousand United States Dollars, this in fact counts for 75.1% of the data.

We can take a quick look at the descriptive statistics of the continuous data:

	age	fnlwgt	educational_num	capital_gain	capital_loss	hours_per_week
count	30161.000000	3.016100e+04	30161.000000	30161.000000	30161.000000	30161.000000
mean	38.438115	1.897992e+05	10.121316	1092.044064	88.302311	40.931269
std	13.134830	1.056506e+05	2.550037	7406.466611	404.121321	11.980182
min	17.000000	1.376900e+04	1.000000	0.000000	0.000000	1.000000
25%	28.000000	1.176280e+05	9.000000	0.000000	0.000000	40.000000
50%	37.000000	1.784290e+05	10.000000	0.000000	0.000000	40.000000
75%	47.000000	2.376300e+05	13.000000	0.000000	0.000000	45.000000
max	90.000000	1.484705e+06	16.000000	99999.000000	4356.000000	99.000000

Figure 3: Descriptive statistics of the continuous variables in the dataset.

We notice that the variable *fnlwgt* (final weight) seems to have a broad range. The final weight variable is described in the dataset description as having to do with the survey weight on the Current Population Survey. For these reasons we chose not to include the *fnlwgt* variable as a predictor variable.

When examining the categorical data we saw that *native-country* had a unique category of "Holand-Netherlands", so this was removed from the dataset.

The following figure is a heat map to visualize the correlation amongst all the variables:

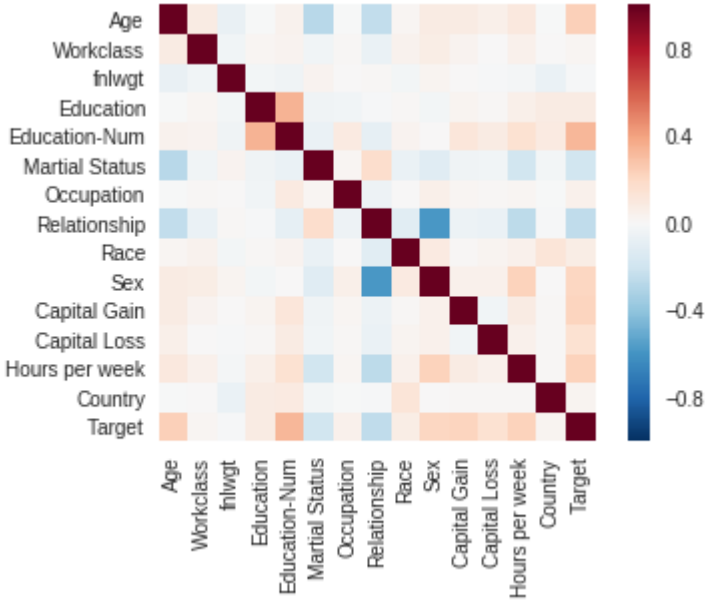

Figure 4: Heat map of original variables

We see a strong, but obvious, correlation between relationship and sex. This is obviously attributed to the fact that traditionally males are husbands and females are wives. We also see strong positive correlation between the variable *Education-num* and the class (Target) and also the variable *Age* and the class (Target). These may be good predictor variables to consider. Lastly we see a correlation between *Education* and *Education-num*. Let's exam this further.

Looking side-by-side of *Education* and *Education-num* we get:

	education	educational_num
0	Bachelors	13
1	Bachelors	13
2	HS-grad	9
3	11th	7
4	Bachelors	13
5	Masters	14
6	9th	5
7	HS-grad	9
8	Masters	14
9	Bachelors	13
10	Some-college	10
11	Bachelors	13
12	Bachelors	13
13	Assoc-acdm	12
14	7th-8th	4

From looking at the table we see that *Education* and *Education-num* are in fact the same variable but *Education-num* is a continuous version of *Education*. We chose to only continue with the variable of *Education-num*.

2.1 Examining Categorical Variables and Their Graphs

We can also examine a few of the categorical variables based on their classes to get a better grasp on the distribution. Here we choose variables which don't contain *too* many categories for ease of visualizing. Each histogram shows the percentage of each category dependent on which class it is in (i.e. either less than or equal \$50K or greater than \$50K).

From *Figure 5* we see that this graph supports claims from our background investigation in section 1, that those of Asian descent (followed by white's) have a higher ratio of high income earnings.

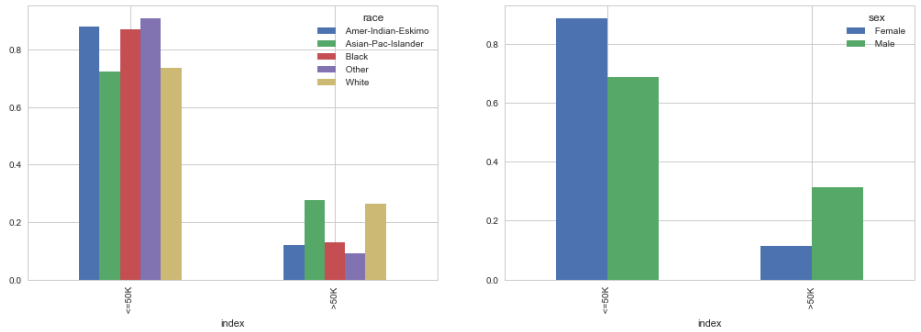

Figure 5: Race and Sex variables

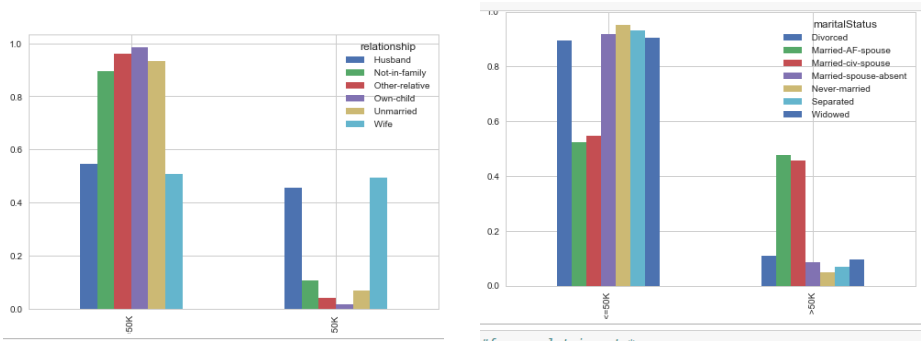

Figure 6: Relationship and Martial Status variables

Examining *Figure 5* also lends support to our sociological information that there is a gender gap when it comes to wages for men and women. However, this graph may be misleading since this chart has no way of telling if there was an even sampling across similar occupations of the males and females in this dataset.

Observing *Figure 6* we see that there is a much larger ratio in those making more than $50K with a relationship status of "husband" or "wife." We also see a similar outcome in the marital-status attribute; that is, individuals that are married have a much larger ratio in the greater than $50K class. We will later see that this plays a very important role.

By looking closer at the feature Martial-Status, we can see that there are actually just two categories hidden; namely, "married" and "not married." So, we decided to adjust the data into these two categories.

	age	workclass	educational_num	marital_status	occupation	relationship	race	gender	capital_gain	capital_loss	hours_per_week
0	39	State-gov	13	not married	Adm-clerical	Not-in-family	White	Male	2174	0	40
1	50	Self-emp-not-inc	13	married	Exec-managerial	Husband	White	Male	0	0	13
2	38	Private	9	not married	Handlers-cleaners	Not-in-family	White	Male	0	0	40
3	53	Private	7	married	Handlers-cleaners	Husband	Black	Male	0	0	40
4	28	Private	13	married	Prof-specialty	Wife	Black	Female	0	0	40
5	37	Private	14	married	Exec-managerial	Wife	White	Female	0	0	40

Once the marital-status was converted we decided to start the real work. In order to use the scikit-learn classifiers however, our categorical data needed to mapped to numerical data. We can see a sample window of the new data in *Figure 7*.

	age	workclass	educational_num	marital_status	occupation	relationship	race	gender	capital_gain	capital_loss	hours_per_week	native_country
0	39	5	13	1	0	1	4	1	2174	0	40	37
1	50	4	13	0	3	0	4	1	0	0	13	37
2	38	2	9	1	5	1	4	1	0	0	40	37
3	53	2	7	0	5	0	2	1	0	0	40	37
4	28	2	13	0	9	5	2	0	0	0	40	4

Figure 7: Sample of cleaned data set with all variables transformed to have numerical values.

2.2 Feature Extraction

We now wish to extract the most relevant features from the dataset. We do this by first looking at which variables are highly correlated. Although correlation does not imply causation this can sometimes be reasonable approach.

To account for the difference in categorical and continuous variables we must be careful. This is not as simple as just using pearsons r correlation coefficient (aside from not all data being continuous) because we made no assumptions about the variables distributions.

To compensate for this we used the *pointbiserialr* function from *scipy* for our continuous data and the *spearmansr* function for our categorical data. The function *pointbiserialr* is used to measure the relationship between a binary variable (our class outcome) and a continuous variable. Similar with other correlation coefficients, *pointbiserialr* varies between -1 and $+1$ (implying a determinative relationship) with 0 implying no correlation.

parameter	abs_corr	correlation
marital_status	0.436133	-0.436133
educational_num	0.335287	0.335287
relationship	0.250998	-0.250998
age	0.241991	0.241991
hours_per_week	0.229480	0.229480
capital_gain	0.221195	0.221195
gender	0.216680	0.216680
capital_loss	0.150222	0.150222
race	0.071666	0.071666
occupation	0.051577	0.051577
native_country	0.023515	0.023515
workclass	0.018040	0.018040

Figure 8: Correlation Table: Variables correlation with target

We decided to take the top seven variables ranked by magnitude of correlation. These [assumed] top 7 predictor variables in order were: Martial-status, Education-number, Relationship, Age, Hours worked per week, Capital-gain, and Gender.

Again we look at a correlation heat-map (*Figure 9.*), but this time we only include our variables that we will use in our model.

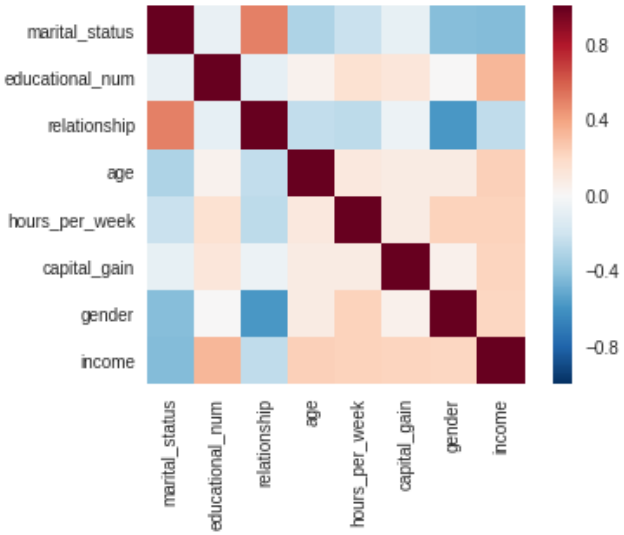

Figure 9: Heat-map of seven featured variables.

We also give a visual of the separation of the classes by using Principal Component Analysis and graphing the outcome.

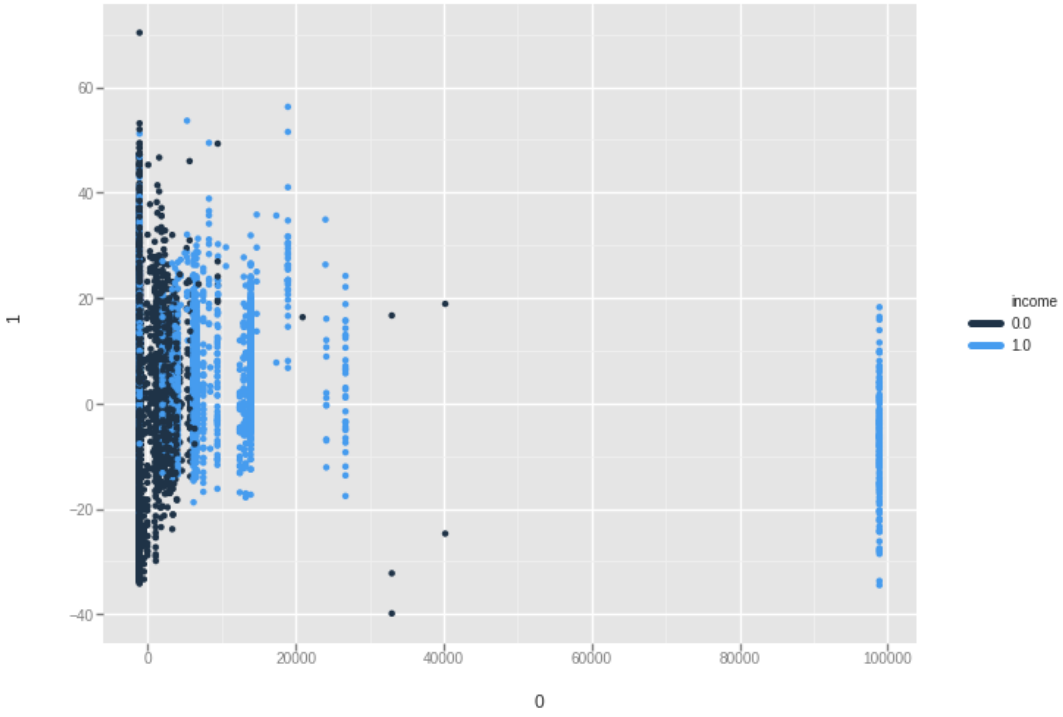

Figure 10: PCA of using the [assumed] top 7 features.

3 Classification Processing

Since we were given a data test set to test our model on, a hold-out of data was not required. In order to make use of all of our available data we used a k-fold cross validation, with $k = 5$. This split the data into five pieces. Four of the sets were used for training and the other was used for testing. This was repeated five times, each time a different set of data was used as the testing set and the other four as the training sets.

We took the mean and standard deviation for these samples, using the score of accuracy, and computed the following table and figure.

Algorithm Name	Mean/Average	Standard Deviation
Linear Regression	83.4024%	(0.004113)
Linear Discriminant Analysis:	82.3149%	(0.004712)
5-NN:	82.6962%	(0.003687)
1-NN:	80.0106%	(0.001232)
9-NN:	83.2333%	(0.002199)
15-NN:	83.6246%	(0.002057)
CART:	81.8076%	(0.005661)
Naive Bayes:	79.3906%	(0.002511)
Decision Tree w/gini:	81.8441%	(0.005537)
Decision Tree:	81.8176%	(0.004981)
Random Forest:	82.5072%	(0.005608)
Support Vector Machine:	82.7890%	(0.003217)

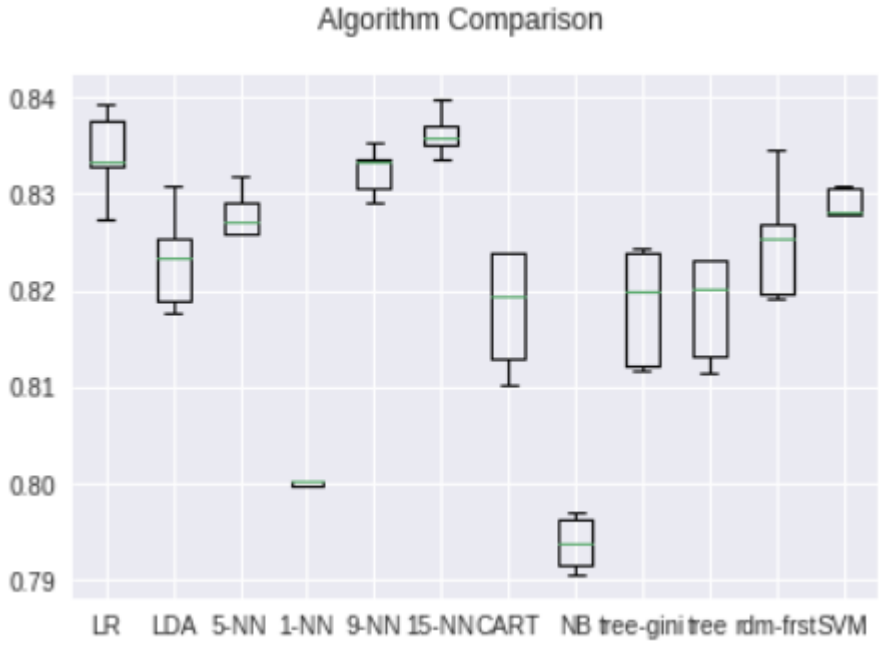

Figure 11: Box and whisker plot comparison of algorithms on data set. These are the averages for the 5-fold cross validation.

Simply choosing to test the dataset on the model with the highest accuracy score from the 5-fold cross-validation we chose to use the 15-Nearest Neighbors models.

4 Experiment (test set)

The supplied test set had 15,050 rows after deleting rows with missing values. The data was then transformed the same was as the data set, that is: marital-status was transformed into one of "married" or "not married", and all categorical data was converted to numerical data (using the same method that was used on the training dataset).

Our best result was the 15-Nearest Neighbor Classifier with an accuracy score of 85.21%. We do take note that SVM had a very small standard deviation, but it took a substantially longer time than the other algorithms to complete.

The confusion matrix produced:

		Prediction outcome		
		p	**n**	**total**
actual value	**p′**	10666 T+	694 F-	11360
	n′	1534 F+	2166 T-	3700
	total	12200	2860	

The following classification report was also produced:

```
Classification report:
              precision    recall  f1-score   support

          0       0.87      0.94      0.91     11360
          1       0.76      0.59      0.66      3700

avg / total       0.85      0.85      0.85     15060
```

5 Conclusion

For future testing we want to consider stratified sampling so that the model is trained on a class distribution that is more like the dataset (and likewise probably similar to reality). Doing this stratified sampling may also increase the *recall* which was not very high.

However, the accuracy score of 85.21% is similar to many other models that reference the dataset. The ics.uci website had a highest accuracy model posted as 85.9%.

We can also see form *Figure 12* that our results were similar to those in other papers.

Dataset size	Testing Correctness % Running Time *Sec.*						
(Training, Testing)	Method						
n = no. of attributes	PSVM	LSVM	SSVM	SOR	SMO	SVM^{light}	RLP
(1605, 30957)	84.00	**84.27**	**84.27**	84.06	-	84.25	78.68
$n = 123$	**0.3**	3.3	1.9	**0.3**	0.4	5.4	9.9
(2265, 30297)	84.13	**84.66**	84.57	84.24	-	84.43	77.19
$n = 123$	**0.5**	5.0	2.8	1.2	0.9	10.8	19.12
(3185, 29377)	84.25	84.55	**84.63**	84.23	-	84.40	77.83
$n = 123$	**0.7**	8.1	3.9	1.4	1.8	21.0	80.1
(4781, 27781)	84.35	**84.55**	**84.55**	84.28	-	84.47	79.15
$n = 123$	**1.2**	8.1	6.0	1.6	3.6	43.2	88.6
(6414, 26148)	84.49	**84.68**	84.60	84.30	-	84.43	71.85
$n = 123$	**1.6**	18.8	8.1	4.1	5.5	87.6	218.8
(11221, 21341)	84.48	**84.84**	84.79	84.37	-	84.68	60.00
$n = 123$	**2.5**	38.9	14.1	18.8	17.0	306.6	449.2
(16101, 16461)	84.78	**85.01**	84.96	84.62	-	84.83	72.52
$n = 123$	**3.7**	60.5	21.5	24.8	35.3	667.2	632.6
(22697, 9865)	85.16	**85.35**	**85.35**	85.06	-	85.17	77.43
$n = 123$	**5.2**	92.0	29.0	31.3	85.7	1425.6	991.9
(32562, 16282)	84.56	**85.05**	85.02	84.96	-	**85.05**	83.25
$n = 123$	**7.4**	140.9	44.5	83.9	163.6	2184.0	1561.1

Figure 12: Testing set correctness and running times on the larger Adult dataset obtained by seven different methods using a linear classifier. Timing comparisons are approximate because of the different machines used, but they do indicate that PSVM has a distinct edge, e.g. solving the largest problem in 7.4 seconds, much faster than any othe r method. Best results are shown in bold. (http://pages.cs.wisc.edu/˜gfung/Glenn_fung_Phd_thesis.pdf)

6 Contributions, Learning Outcomes, and Difficulties

6.1 Contributions

While the project was split between Nathalie and I we both had very different contributions. While Nathalie did majority of the the pre-processing and data clean up, along with some of the earlier exploration and graphs, **I made the contributions of**:

- solely finding proper functions for finding correlation and doing feature-extraction
- solely making the visualizations of PCA and Heat-Maps

- K-fold Cross Validation
- Machine-learning algorithm testing and comparison
- testing the dataset on our chosen model
- solely writing this entire LaTeXreport [redacted] *blacked out from public

6.2 Learning Outcomes

Fortunately I have had previous experience using jupyter-notebooks, python, and scikit-learn on a similar type of project to this one. However, doing this project and making the presentation forced me to understand exactly how the functions worked and how to apply the theory. For example, I knew of k-fold cross-validation being a method for training a model before this project, but I didn't know how cross-validation worked and it's advantages. I learned that this method works better than a traditional hold-out method because all of the data from the training dataset is used to obtain the model.

Doing this project I also learned how to read and interpret a confusion matrix. Another technical skill that I learned is also how to make a heat-map to visualize the correlation between variables.

One of the most interesting things I learned is discussed in the following subsection.

6.3 Difficulties Faced

The biggest difficulty that I faced when doing this project was dealing with the categorical variables; I've never used the library scikit-learn to make models that had categorical variables before this. One really interesting function I came across doing this project was the *scipy* library function *pointbiserialr* which gave a correlation coefficient for continuous and a binary variable. Before this, I didn't know that this type of correlation could be measured.

Given that I had more time I would've finished the code that I started for testing the stratified sampling for cross-validation to see if the results would positively affect the recall score.

[redacted] **(blacked out from public)