PREDICT SAVINGS FROM CENSUS DATA

GOAL

Generate classifier to predict the income level for the person represented by a record. Incomes have been binned at the $50K level to present a binary classification problem.

SUMMARY

US Census data provides anonymous information on each profile such as age, education, wages per hour, etc. Of these features, seven are continuous while the rest are nominal. Training set provided is used to generate classifiers – Logistic Regression, Decision Tree and Random Forest were chosen for this project. Using five-fold cross validation to evaluate the classifiers, it was determined that Random Forest gave the best score. Since response variable is binary, metrics for evaluation is F-score.

The best value for Random forest parameter 'maximum depth' was chosen – it was 23. The modified Random forest model was used to predict F1-score on both the training data and test data. Although training F1 score saw 10% improvement, the F1 score on test data is at 0.45. It may be worthwhile to try Gradient Boosting classifier to check for improvements in prediction. This is not covered in this project at present.

To gain insights on which features have positive effect on savings – few such features are Age, Dividends, more family members working, more weeks worked. Males appear to have more savings. Also, at least a high school graduation has more savings. More insights are discussed in the following section.

INSIGHTS

Continuous Variables

From entire dataset including all profiles, we get the following statistics on continuous variables.

	count	mean	std	min	25%	50%	75%	max
Variables								
age	199523.0	34.494199	22.310895	0.0	15.0	33.0	50.0	90.0
wage_per_hour	199523.0	55.426908	274.896454	0.0	0.0	0.0	0.0	9999.0
cap_gains	199523.0	434.718990	4697.531280	0.0	0.0	0.0	0.0	99999.0
cap_loss	199523.0	37.313788	271.896428	0.0	0.0	0.0	0.0	4608.0
dividends	199523.0	197.529533	1984.163658	0.0	0.0	0.0	0.0	99999.0
num_worked_for_employer	199523.0	1.956180	2.365126	0.0	0.0	1.0	4.0	6.0
weeks_worked	199523.0	23.174897	24.411488	0.0	0.0	8.0	52.0	52.0

PREDICT SAVINGS FROM CENSUS DATA

When considering only the profiles that have savings > 50,000, the metrics on the continuous variables is :

	count	mean	std	min	25%	50%	75%	max
Variables								
age	12382.0	46.266193	11.830906	16.0	38.0	45.0	53.0	90.0
wage_per_hour	12382.0	81.640284	431.364773	0.0	0.0	0.0	0.0	9999.0
cap_gains	12382.0	4830.930060	16887.627002	0.0	0.0	0.0	0.0	99999.0
cap_loss	12382.0	193.139557	607.542507	0.0	0.0	0.0	0.0	3683.0
dividends	12382.0	1553.448070	6998.071762	0.0	0.0	0.0	363.0	99999.0
num_worked_for_employer	12382.0	4.003715	2.118183	0.0	2.0	4.0	6.0	6.0
weeks_worked	12382.0	48.069617	12.259412	0.0	52.0	52.0	52.0	52.0

Comparing the above two tables, note that:
1. Median age is higher at 45 when higher savings compared to 33. It is intuitive that people have more savings as they grow older.
2. Among people with savings > 50,000, 25% have Dividends > 363 units. This is also intuitive that as dividends increase one gets more savings.
3. When more family members are working , savings are higher. 'num_worked_for_employer' has higher 25%, median, 75% when savings are higher.
4. 75% of people have worked the full year when savings are higher. This is intuitive too.

Nominal Variables

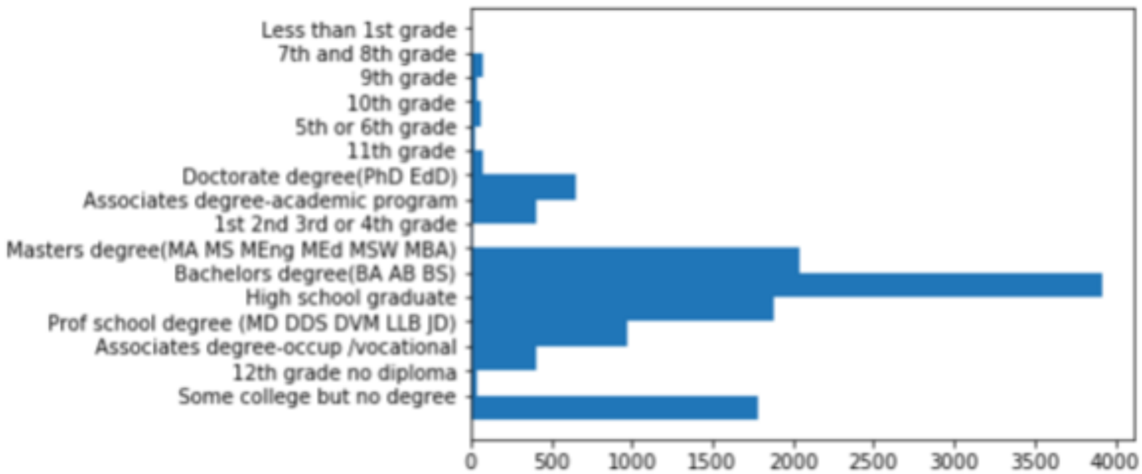
Nominal variables are now evaluated for their effects on savings > 50,000. First, a quick overview of the top category within each feature that affects higher savings is determined. A snapshot of few such results:

Nominal Feature	Top Category
worker_class	Private
industry_cd	45
occupation_cd	2
education	Bachelors degree(BA AB BS)
enrolled	Not in universe
marital_status	Married-civilian spouse present
major_industry_cd	Manufacturing-durable goods
major_occupation_cd	Executive admin and managerial
race	White

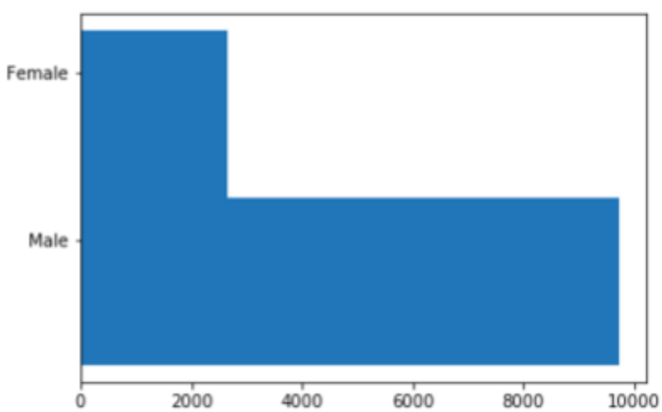
PREDICT SAVINGS FROM CENSUS DATA

Looking in more detail into few select features, following insights were obtained.

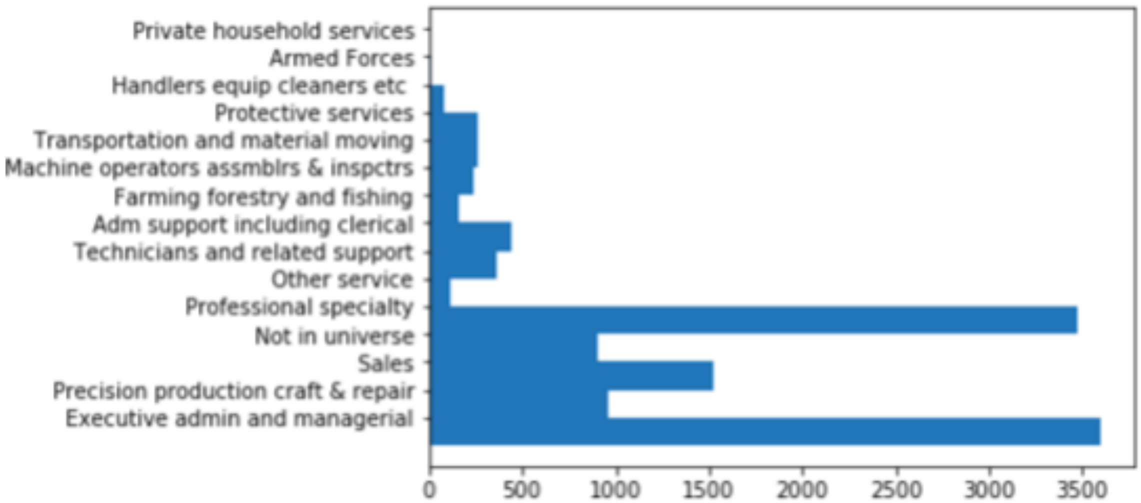
1. Education: top 3 education levels among people with higher savings is ('Bachelors degree(BA AB BS)', 3915), (' Masters degree(MA MS MEng MEd MSW MBA)', 2038), (' High school graduate', 1879)]. So minimum High school graduation appears to affect the savings.



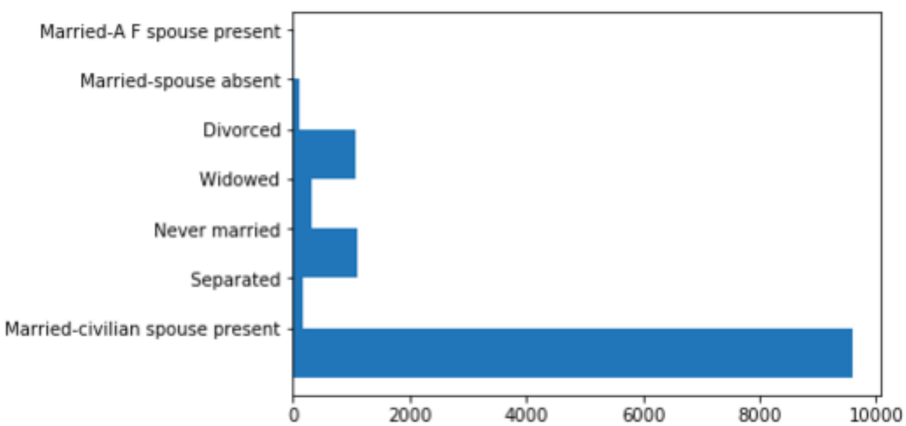
2. Sex: Males have higher savings than females. [(' Male', 9719), (' Female', 2663)]



3. Major occupation code: The top occupations with the highest savings are [('Executive admin and managerial', 3593), (' Professional specialty', 3475), (' Sales', 1524)].



4. Marital Status: It appears being married has significant positive effect on savings. [(' Married-civilian spouse present', 9600), (' Never married', 1117), (' Divorced', 1066)]



Further investigation on other nominal parameters could provide more insights.

CHALLENGES

+ Finding illogical / irrelevant entries in each column, and handling then during analysis. It is considered a unique category which in fact it is not.
+ When finding top category, it was difficult getting code to bypass all " ?"
+ Linear Regression and Gradient Boosting classifiers take a long time to build especially when using k-fold cross validation. Hence, couldn't include Gradient Boosting classifier in this project