CENSUS INCOME PREDICTION MEMORANDUM

DATE: July 5, 2018
TO: ****
FROM: Homer Kay, Data Scientist
SUBJECT: Analysis and recommendation for production machine learning.

Purpose

The purpose of this analysis is to predict an income target of greater or less than $50,000 based on a number of features. Exploratory data analysis, application of machine learning models, and performance evaluation are included in this analysis.

Data Cleaning

Data cleaning included: combining the originally split test/train data sets, removing null and duplicate features, defining numeric equivalents for categorical features, and scaling data for the appropriate models.

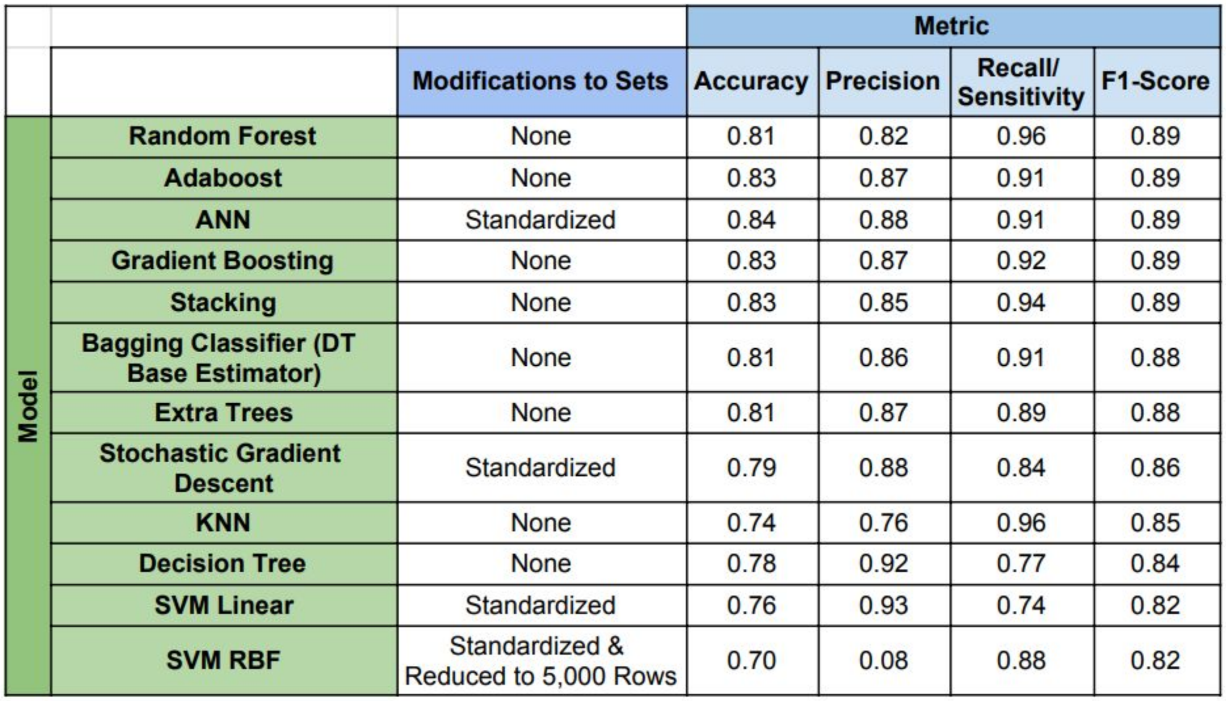
Exploratory Data Analysis

Evaluating features independently revealed the following results. The weight feature is well distributed with a median around 180,000. Education feature shows clear peaks at high school grads, some college complete, and bachelors degree. The vast majority input working near 40 hours/week. The age feature is well distributed starting at age 18 and with a median of ~32 years. Over 85% of replies were from Caucasian persons with the rest of majority being African American. 66% of replies were from males. A wide variety of professions are represented being services, sales, executives, specialty professionals to name a few. In the future census I recommend that professions should be limited to fewer selections; there are to many vague options in the current system for clear feature definition. The marriage selection could also be grouped as separated, never-married, and married. 73% of the workclass responses were from the private sector. As would be expected, over 90% of the responses were native's of the United States.

Predictive Models

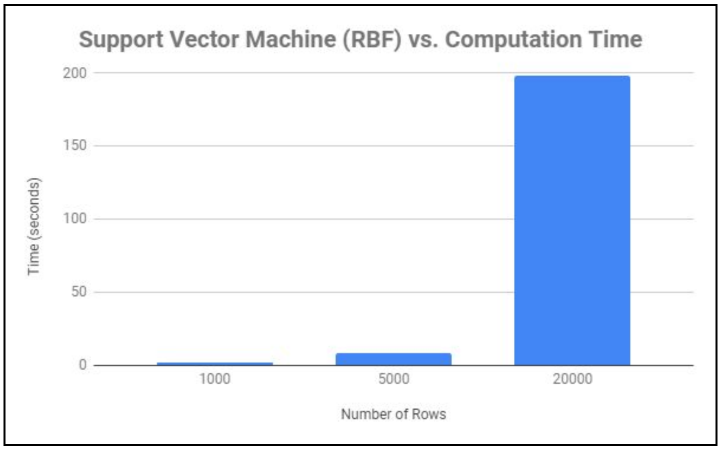
The training/test was split 70/30; for the size of the data (45,000+ rows) this split is appropriate. On some occasions the data sets were normalized and/or reduced for model necessity or computational efficiency. The results can be seen below.

Results



			Metric			
		Modifications to Sets	Accuracy	Precision	Recall/ Sensitivity	F1-Score
Model	Random Forest	None	0.81	0.82	0.96	0.89
	Adaboost	None	0.83	0.87	0.91	0.89
	ANN	Standardized	0.84	0.88	0.91	0.89
	Gradient Boosting	None	0.83	0.87	0.92	0.89
	Stacking	None	0.83	0.85	0.94	0.89
	Bagging Classifier (DT Base Estimator)	None	0.81	0.86	0.91	0.88
	Extra Trees	None	0.81	0.87	0.89	0.88
	Stochastic Gradient Descent	Standardized	0.79	0.88	0.84	0.86
	KNN	None	0.74	0.76	0.96	0.85
	Decision Tree	None	0.78	0.92	0.77	0.84
	SVM Linear	Standardized	0.76	0.93	0.74	0.82
	SVM RBF	Standardized & Reduced to 5,000 Rows	0.70	0.08	0.88	0.82

Models are ranked in descending order based on their F1-Score (Harmonic mean of Recall and Precision). I found tuning boosting and bagging methods to increase model accuracy significantly (5%+) compared to tuning base learners, KNN, Decision Tree, etc. In contrast to their lack of performance the base learners are critical when comparing and validating the performance of more complex ensemble models. One interesting note is the potential for SVM RBF, but the lack of computational resources makes it hard to predict with the full dataset. In the below figure one can see the exponential increase in computation time with the number of rows, hence the justification for the reduced number of 5,000 rows in the analysis.



Conclusion and Recommendations

The top five models have the same F1-Score; furthermore the ANN model has the highest accuracy and precision, while Random Forest has the highest Recall. Upon further analysis of their confusion matrices and without a specific business case to prefer (bias) true positives or negatives (as in a medical tumor case) I believe the strongest model for the application to be ANN. Considering the fact that naturally the model will give a 75% chance of <50K (this is because 75% of the replies are below that income), an increase to 84% is significant.

ANN is superior to the Random Forest in that it gets more True Positives (>50K), which because of the 75% bias towards <50K is an attractive trait. Moreover the ANN model has more even False Negatives and False Positives, making it more centered around the dataset.

With the highest Accuracy and F1-Score I recommend we place the ANN model into production. This model, as well as all others, has been cross validated to confirm that all the variation in the dataset has been appropriately captured in while training the model.