

CS7641: Machine Learning

Assignment 1 Analysis Report

Jun Zhang

September 22, 2019

1 Introduction

Banking crisis is always an interesting topic to study. It is important but difficult to be predicted. Traditionally, financial analysts apply economic theories like business cycle to foresee a crisis. Machine learning techniques, on the other hand, can provide a new way for people to understand the dynamics of a market.

Online shopping is close to everyday life. It is easy to understand and easy to be implemented. Also, knowing what features would impact on online shopping behavior help merchandisers and website designers to read their customers better. It would also have positive impact on shopping experience online.

Above two topics seem to be very different in context, but they share commons in terms of datasets. First, both datasets are large enough to apply machine learning techniques. Second, their target outputs are boolean values. They both need to answer questions like: would crises happen based on given conditions? Or would shopper make online order based on their behaviors? Third, their outputs are determined by a group of features. Some of them are categorical(digits or strings), while some of them are continuous. Last, the datasets are kind of imbalanced, which means the possibility of crises or online shopping is low at around 10%-15%.








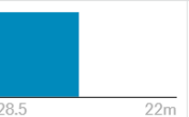



2 Description of Datasets

African Crises dataset and Online Shopping Intention dataset are both available to be downloaded at www.kaggle.com.

2.1 African Crises Dataset

The first dataset is derived from Reinhart et. al's Global Financial Stability dataset. It presents the Banking, Debt, Financial, Inflation and Systemic Crises that occurred in 13

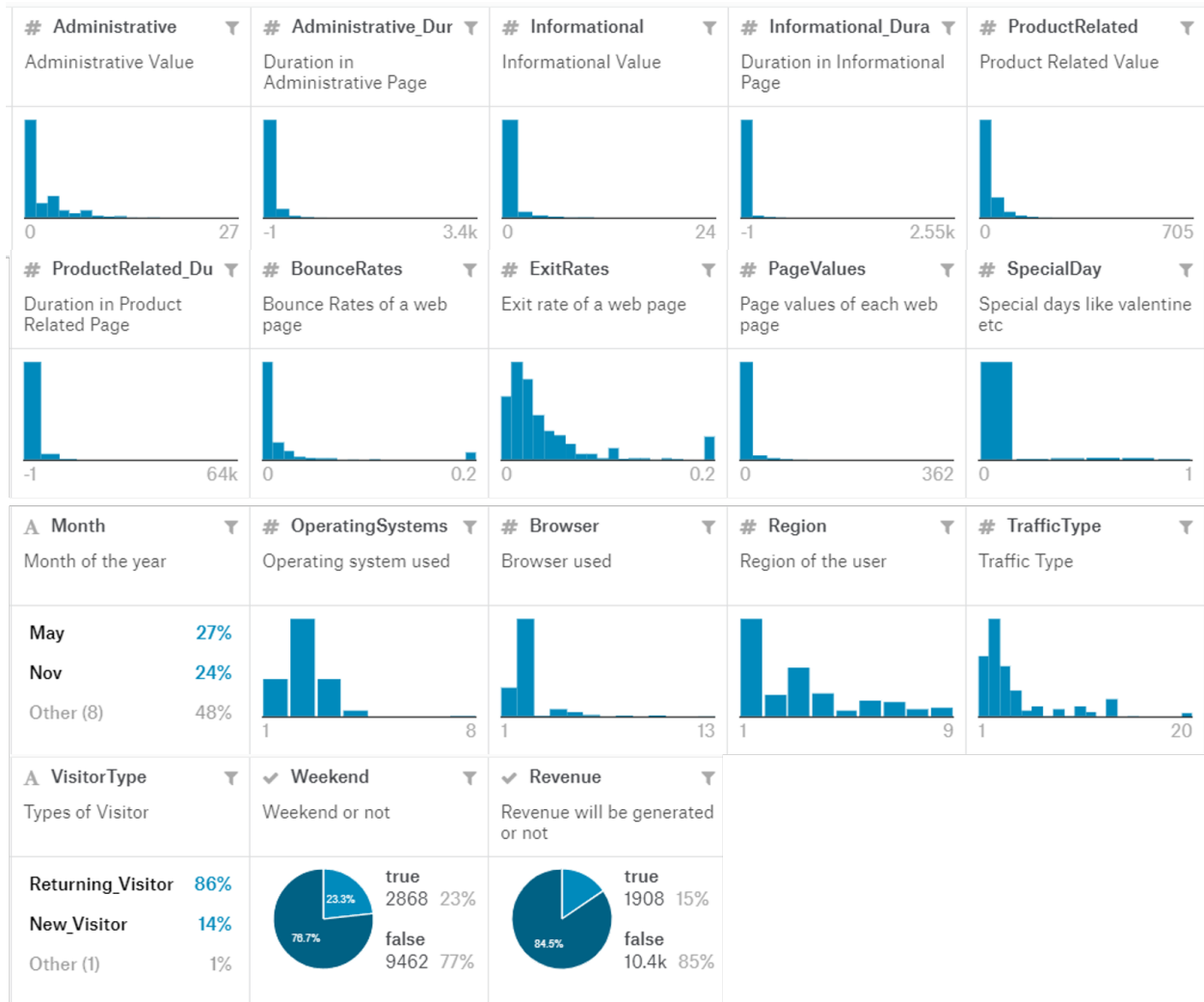
African countries from 1860 to 2014: Algeria, Angola, Central African Republic, Ivory Coast, Egypt, Kenya, Mauritius, Morocco, Nigeria, South Africa, Tunisia, Zambia and Zimbabwe. There are 1059 data points and 13 features that could affect the prediction of a crisis. Most features are around a level of 0. features like systemic_crisis and currency_crisis are expected to be closely related to banking_crisis(the target).

# case ▼ A number which denotes a specific country 	A cc3 ▼ A three letter country code EGY 15% ZAF 11% Other (11) 75%	A country ▼ The name of the country Egypt 15% South Africa 11% Other (11) 75%	# year ▼ The year of the observation 	# systemic_crisis ▼ "0" means that no systemic crisis occurred in the year and "1" means that a systemic crisis occurred in the year. 	# exch_usd ▼ The exchange rate of the country vis-a-vis the USD 
# domestic_debt_in_d ▼ "0" means that no sovereign domestic debt default occurred in the year and "1" means that a sovereign domestic debt 	# sovereign_external ▼ "0" means that no sovereign external debt default occurred in the year and "1" means that a sovereign external debt 	# gdp_weighted_defa ▼ The total debt in default vis-a-vis the GDP 	# inflation_annual_cp ▼ The annual CPI Inflation rate 	# independence ▼ "0" means "no independence" and "1" means "independence" 	# currency_crises ▼ "0" means that no currency crisis occurred in the year and "1" means that a currency crisis occurred in the year 
# inflation_crises ▼ "0" means that no inflation crisis occurred in the year and "1" means that an inflation crisis occurred in the year 	A banking_crisis ▼ "no_crisis" means that no banking crisis occurred in the year and "crisis" means that a banking crisis occurred in the year no_crisis 91% crisis 9%				

This dataset has no missing values, so I keep the full set of data. To help better predict crises, I classified "exch_usd" into 4 categories based on a range of values(0:exch_usd< 1, 1:exch_usd< 10, 2:exch_usd< 100, 3:exch_usd≥ 100). I also classified "inflation_annual_cpi" into 4 categories(0:exch_usd< 0, 1:exch_usd< 4, 2:exch_usd< 10, 3:exch_usd≥ 10).

2.2 Online Shopping Intention Dataset

There are 12330 data points and 17 features that would affect final target "Revenue". Features like "ProductRelated", "SpecialDay", "Month" and "Weekend" are expected to be closely related to online shopping.



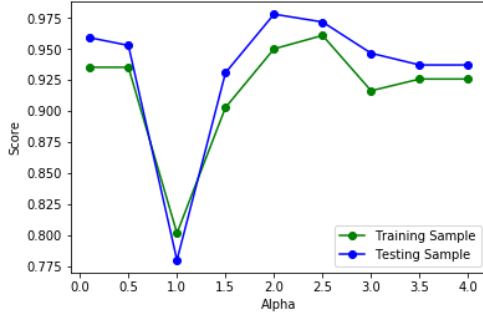
In this dataset, there are 14 points missing values for features related to "Administrative" and "Informational". I remove those 14 points to create a null missing dataset. I also convert string values to integers(e.g. True=1, False=0; Feb=2, Dec=12, etc.). No other special handling applied to this dataset.

3 Learning Algorithms Used

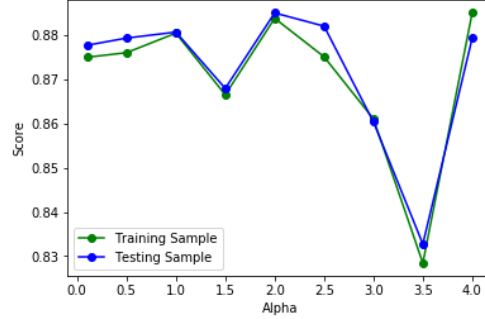
In this section, I am going to discuss how well 5 machine learning algorithms(Neural Network, Support Vector Machine, etc) performed on my two data sets. All algorithms used are available in Python Sklearn package. I applied 10-fold cross validation for all algorithms except for SVM (it took too much time to run a 10-fold CV).

3.1 Neural Network

In this section, I used the Multi-Level Perceptron classifier in Sklearn. For hidden levels, I fixed to (100,100) and also set minimum leaves number to 3. I grid searched a list of alpha (0.1, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4) and plot corresponding accuracy scores. The alpha level that makes both training and testing sets achieve the best(or relatively better) accuracy score is chosen for this model, which is 2 for both datasets. Running time for a singly Neural Network model is 0.16min and 0.29min respectively.

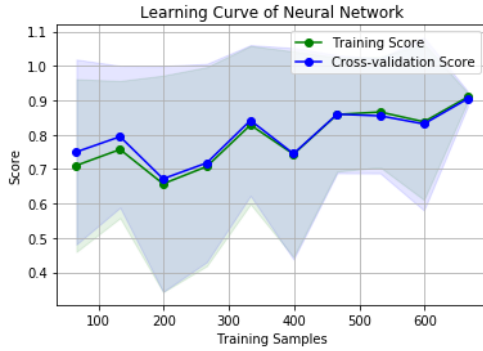


(a) African Crises

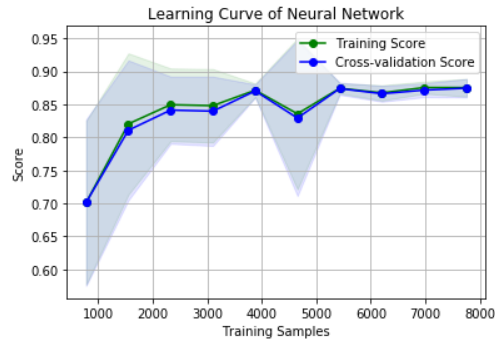


(b) Shopping Intention

From the graphs of learning curve, we can see that as data size grows, selected models performs better.



(a) African Crises

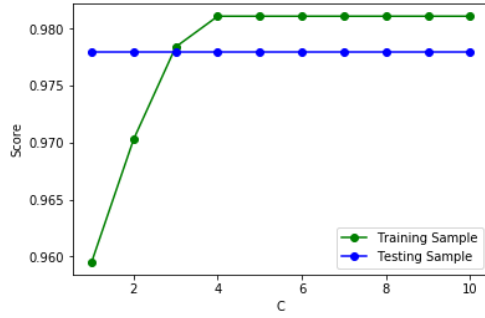


(b) Shopping Intention

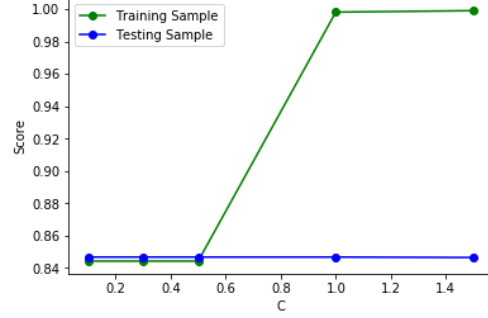
3.2 Support Vector Machine

In this section, I used the C-Support Vector classifier in Sklearn. I tried to use Linear, RBF and Polynomial kernals. However, for Online Shopping Intention dataset, only RBF kernel can return values within 20 mins. I also grid searched a list of C ([1-9] for African Crises dataset and [0.1, 0.3, 0.5, 1, 1.5] for Online Shopping Intention dataset) and plot corresponding accuracy scores. The C level that makes both training and testing sets achieve the best(or relatively better) accuracy score is chosen for this model, which is 4 for African

Crises dataset and 1 for Online Shopping Intention dataset. Running time for a singly SVM model is 0.3min and 7.54min respectively.

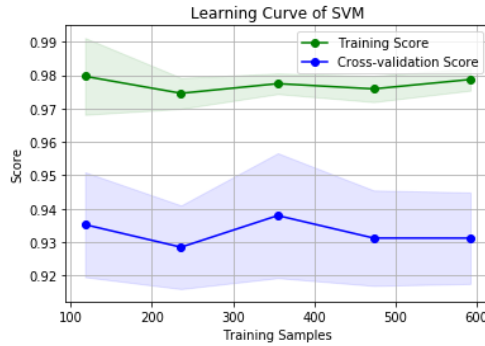


(a) African Crises

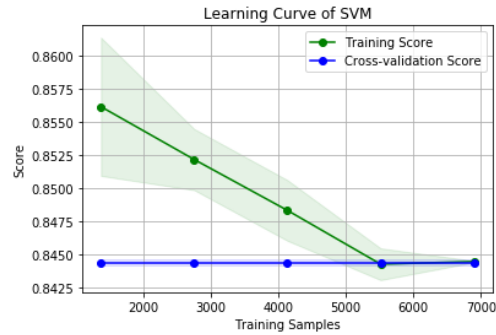


(b) Shopping Intention

The learning curve of first dataset is quite flat while that of second dataset slopes downward. That's to say, the selected model works better for a smaller size(around 1000 data points) of Shopping Intention dataset. Consider the original size of the 2 datasets, we cannot conclude that our model works fine for a larger data size(say, 2000 or 3000 data points) of African Crises dataset.



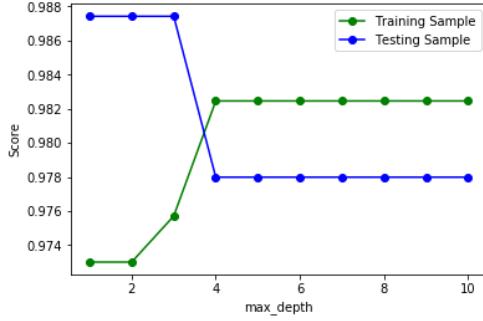
(a) African Crises



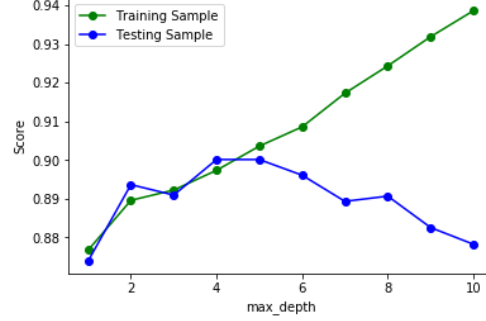
(b) Shopping Intention

3.3 Decision Tree

For Decision Tree Classifier, I grid searched a list of max_depth ([1-10] for both datasets, step=1) and plot corresponding accuracy scores. The max_depth level that makes both training and testing sets achieve the best(or relatively better) accuracy score is chosen for this model, which is 3 for African Crises dataset and 4 for Online Shopping Intention dataset. Running time for a singly Decision Tree model is 0.001min and 0.02min respectively.

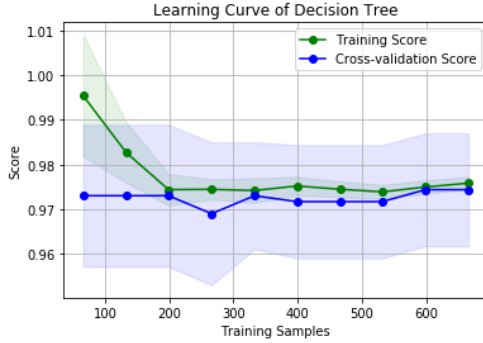


(a) African Crises

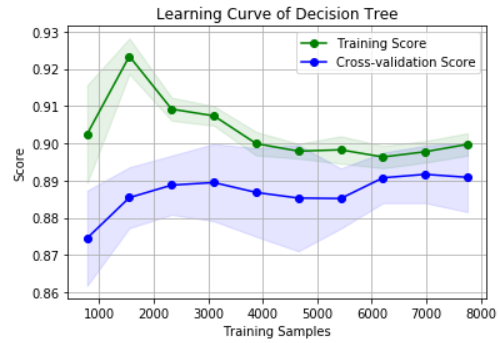


(b) Shopping Intention

From learning curve of African Crises dataset, we can see that training data perform worse when data size increases from less than 100 to 200. Curves becomes stable when size larger than 200. Testing data, however, is not sensitive to size increases. For learning curve of Online Shopping Intention dataset, training data still perform worse when data size increases, even though there is a jump as data size increases from less than 1000 to less than 2000. Testing data performs slightly better as data size increases.



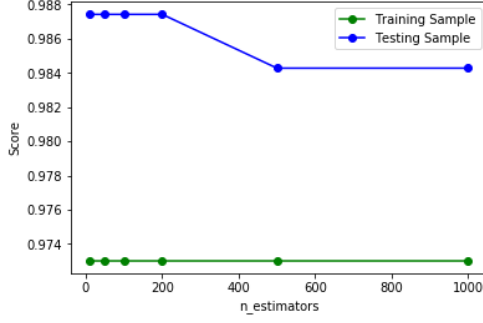
(a) African Crises



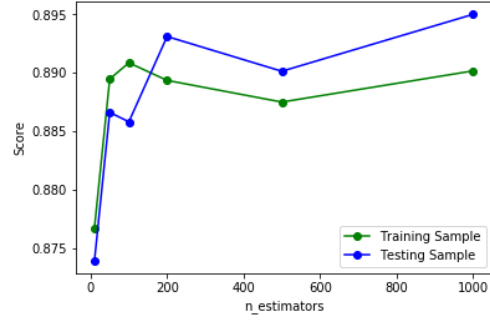
(b) Shopping Intention

3.4 Boosting

In this section, I used the AdaBoost classifier in Sklearn. I grid searched a list of number of weak estimators (10, 50, 100, 200, 500, 1000) and plot corresponding accuracy scores. The number level that makes both training and testing sets achieve the best(or relatively better) accuracy score is chosen for this model, which is 200 for both models (even though $n_estimators=1000$ also looks good for second dataset, it's preferable to choose a smaller n). Running time for a singly AdaBoost model is 0.016min and 1.26min respectively.

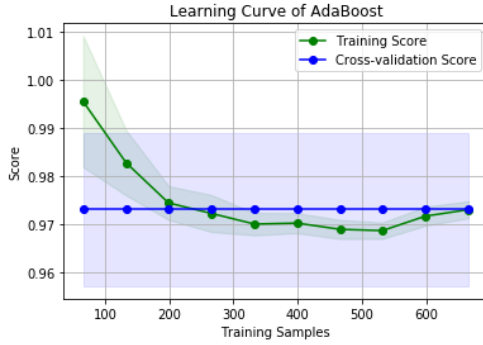


(a) African Crises

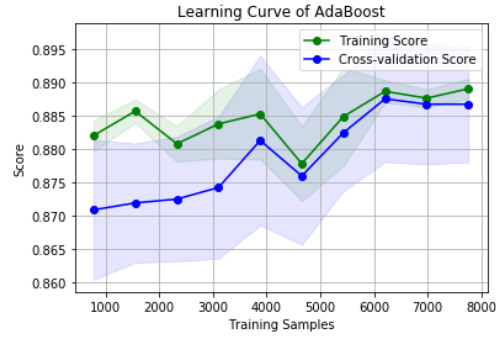


(b) Shopping Intention

From learning curve of African Crises dataset, we can see that training data perform worse when data size increases from less than 100 to more than 300. Curves becomes stable when size larger than 300. Testing data, however, is not sensitive to size increases. For learning curve of Online Shopping Intention dataset, comparing to testing data performance, training data performance is relatively indifferent to size change.



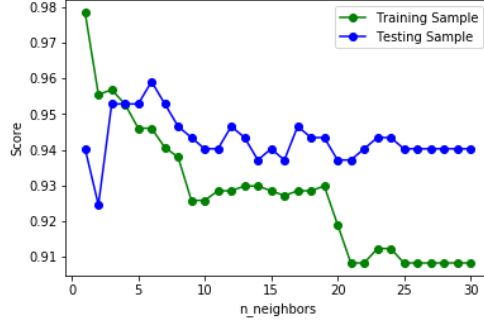
(a) African Crises



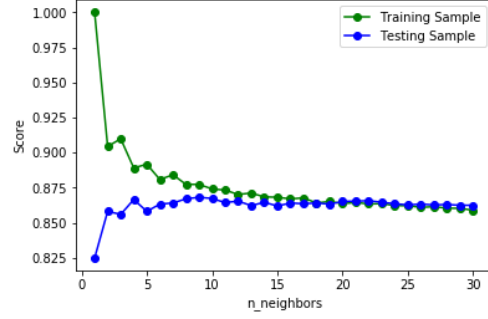
(b) Shopping Intention

3.5 K-Nearest Neighbors

For K-Nearest Neighbors Classifier, I grid searched a list of K ([1-30] for both datasets, step=1) and plot corresponding accuracy scores. The K value level that makes both training and testing sets achieve the best(or relatively better) accuracy score is chosen for this model, which is 6 for African Crises dataset and 9 for Online Shopping Intention dataset. Running time for a singly KNN model is 0.001min and 0.03min respectively.

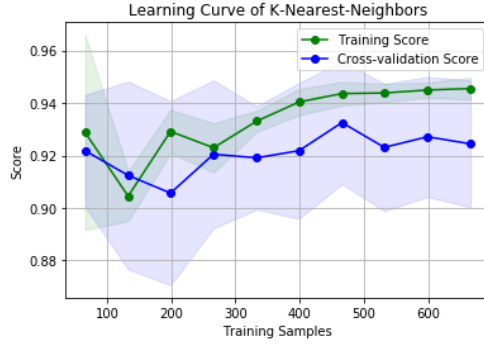


(a) African Crises

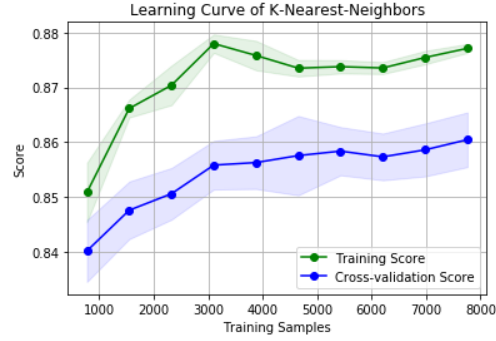


(b) Shopping Intention

From learning curve of African Crises dataset, we can see that training data perform better as data size increases. Testing data, however, is not sensitive to size increases. For learning curve of Online Shopping Intention dataset, both training and testing data perform better as size increases.



(a) African Crises



(b) Shopping Intention

4 Conclusion and Discussion

In this assignment, I grid searched parameters to find a suitable model, and also create learning curves to further validate this model. Based on grid search and learning curve, Neural Network works the best for African Crises dataset and KNN works the best for Online Shopping Intention dataset. However, when I looked at precision rate on confusion tables, I found Decision Tree model works best (for African Crises dataset, it has a highest precision rate(0.91) for Crisis and highest precision rate(0.99) for no_Crisis; For Online Shopping Intention dataset, it has a highest precision rate(0.78) for Buying and highest precision rate(0.91) for no_Buying). Since both datasets are highly skewed, I think resampling with different ratios with higher weight to minority data points would help to achieve one more consistent model for each dataset.