Final Project Report

Xi Yang, 16216573, Group 30

Introduction

One of the most important big data source is economy. Everyday millions of data are produced. Data such as GDP, inflation, stock marketing etc. are every interesting for analysis to extract the useful information for future. All the macro-economy data sets are also closely related with the data sets collected locally. Combining macro-data with local data, a clear picture of distribution of economy based on geo location can be drawn. One of the interesting and easy to obtain data is Local Area Unemployment Statistics data set available from Bureau of Labor Statistics of United States. The unemployment index is always a good indicator of the economy.

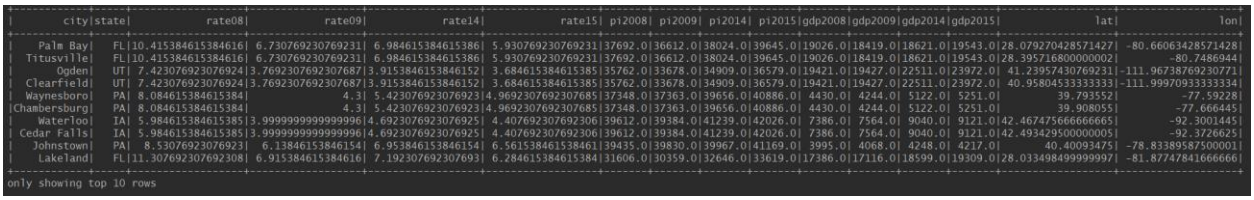
Intuitively, the local unemployment rate should be reflected in the local economic data such as local area GDP and local area per capital income. The ideal relationship would be expected as negatively associated which means the lower the unemployment rate, the higher the GDP and per capital income will be. However, the results must be supported by data analysis. Such types of analysis include data integration, plotting, comparison and model training and predicting. Considering the amount of data that can be collected, scaling analysis method required to process these incoming data especially when we start to talk about data at national or global level. Therefore, the project is designed to target on analysis the relationship among geo location, GDP, personal income and unemployment rate.

Datasets

The datasets I used in this project are multiple csv files from two different sources as Bureau of Labor Statistics of United States and Bureau of Economic analysis of United States, all the datasets are zipped and stored at https://drive.google.com/open?id=1g85KGgpqW1eOKOAqwQEAqbF_pZBVSf-z.

There are three four major data sets. CA1_1969_2015_All_AREAS.csv contains all the GDP information categorized by metropolitan areas of USA; RIP_2008_2015_MSA.csv contains all the personal income information also categorized by metropolitan areas of USA; the la_data_Metro.txt contains all the information about unemployment rate categorized by metropolitan areas of USA; the zip_codes_states.csv contains all the geo location information. Other files are keys that facilitate the connections between each dataset. Also, the unemployment rate county data has been investigated first but due to the lack of county information on personal income, in the final version of the project all the information are associated with metropolitan based data. In this project, data from 2006, 2008, 2009, 2014 and 2015 are extracted and used in the classification analysis, the code can be easily updated to include more years. GDP, personal income and unemployment rates are finally aggregated together by geo location information and a sample of final data frame of first 10 records are shown below. The classification

process is based on this data frame obtained from spark map-reduce-aggregation process.



city	state	rate08	rate09	rate14	rate15	pi2008	pi2009	pi2014	pi2015	gdp2008	gdp2009	gdp2014	gdp2015	lat	lon
Palm Bay	FL	10.415384615384616	6.730769230769231	6.984615384615386	5.930769230769231	37692.0	36612.0	38024.0	39645.0	19026.0	18419.0	18621.0	19543.0	28.079270428571427	-80.660634285714285
Titusville	FL	10.415384615384616	6.730769230769231	6.984615384615386	5.930769230769231	37692.0	36612.0	38024.0	39645.0	19026.0	18419.0	18621.0	19543.0	28.395716800000002	-80.7486944
Ogden	UT	7.423076923076924	3.769230769230768	3.9153846153846152	3.6846153846153845	35762.0	33678.0	34909.0	36579.0	19421.0	19427.0	22511.0	23972.0	41.239574307692311	-111.96738769230771
Clearfield	UT	7.423076923076924	3.769230769230768	3.9153846153846152	3.6846153846153845	35762.0	33678.0	34909.0	36579.0	19421.0	19427.0	22511.0	23972.0	40.958045333333333	-111.99970933333334
Waynesboro	PA	8.084615384615384	4.3	5.4230769230769231	4.9692307692307685	37348.0	37363.0	39656.0	40886.0	4430.0	4244.0	5122.0	5251.0	39.793552	-77.59228
Chambersburg	PA	8.084615384615384	4.3	5.4230769230769231	4.9692307692307685	37348.0	37363.0	39656.0	40886.0	4430.0	4244.0	5122.0	5251.0	39.908055	-77.666445
Waterloo	IA	5.984615384615385	3.9999999999999996	4.6923076923076925	4.4076923076923066	39612.0	39384.0	41239.0	42026.0	7386.0	7564.0	9040.0	9121.0	42.467475666666665	-92.3001445
Cedar Falls	IA	5.984615384615385	3.9999999999999996	4.6923076923076925	4.4076923076923066	39612.0	39384.0	41239.0	42026.0	7386.0	7564.0	9040.0	9121.0	42.493429500000005	-92.3726625
Johnstown	PA	8.530769230769231	6.138461538461154	6.9538461538461154	6.5615384615384615	39435.0	39830.0	39967.0	41169.0	3995.0	4068.0	4248.0	4217.0	40.40093475	-78.833958750000015
Lakeland	FL	11.307692307692308	6.9153846153846161	7.1923076923076931	6.2846153846153841	31606.0	30359.0	32646.0	33619.0	17386.0	17116.0	18599.0	19309.0	28.033498499999997	-81.877478416666661

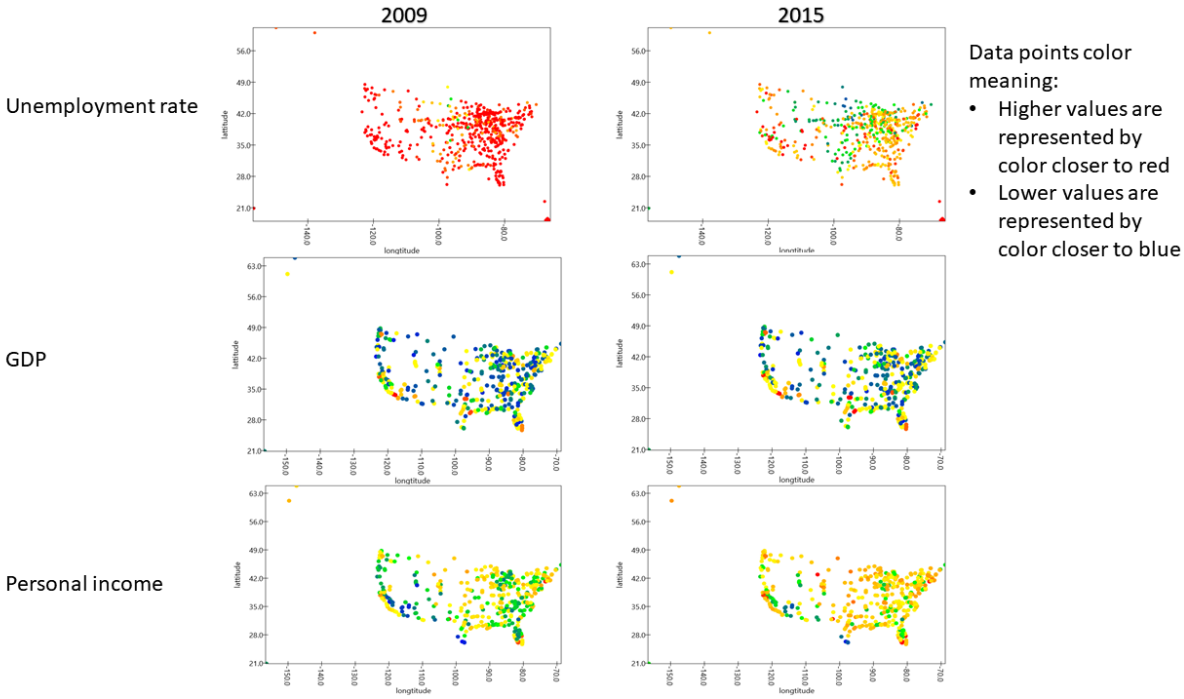
only showing top 10 rows

Approach

To complete the project, we need to finish two tasks. One is data integration and the other is classification. Both parts will be done with Scala using spark frame work which allow easy scale up and can handle large data set up to TB magnitude. To integration the data, the major functions will be used is map-reduce-aggregation. The RDD is the major data structure while data frame will be used later for a better data organizing and format for classification process. The classification models are logistic regression, random forest and multilayer perceptron (neuron networks). All these three classifiers are available from spark.ml package. The yield results will be analyzed not only using accuracy but also precision, recall and f-score will be reported for comparison. In the multi-layer perceptron model, five layers are used with 3 nodes in input, 40, 80, 40 nodes in hidden layers respectively and 40 nodes in the output. The block size is set to 128. In the logistic regression model, the Gaussian logistic model is used since the all the data are in the continuous datatype as double. For the random forest model, the maximum number of trees is set to 10.

Data integration, modeling and results interpretation

The project is divided into two separate parts. The first part is focused on data integration. The major object is to link yearly organized GDP, personal income and unemployment rate data using geo location information. Specifically, all the GDP, personal income and unemployment rate data have at least one attribute that can be directly or indirectly related to the geo location information. The geo location information is represented by either city-state pair or latitude-longitude pair information. The data also is visualized in 2D figures in the first part. The latitude and longitude are used as x, y parameters to illustrate the location distribution and the value of each data point is represented via color. More detail on data point coloring, the highest value is assigned to red and the lowest value is assigned to blue while all other values in between has its own color rendered based on its value relative to highest and lowest. For example, a data point with a value closed to median might be assigned a color as yellowish green. The system produced figures associated with their data sources are depicted as blow, you can refer them by their title.



The goal of the first part is to finally produce a data frame with all geo-location, GDP, personal income and unemployment rate together. After achieving this goal, in the second part, a linear regress model will be first built to test if the GDP, unemployment rate and personal income data have a linear relationship. Later classification models will be tested on the dataset to check if we can use the GDP, unemployment rate and personal income data to prediction the geo-location (city). This will be interesting since different areas have different economy patterns and the question can be answered here is whether the economy patterns can be represented by these three attributes.

First analysis aims to answer the question that whether the GDP, personal income and unemployment rate has a linear relationship. If there is a linear relationship existed, a linear regression model can be built based on the data. Treating GDP and unemployment rate as variables and personal income as outcome, a linear regression model was built with R^2 value equals to 0.05 which indicates that the linear model was not valid. We cannot use the model to predict the personal income value since only about 5% of the data can be explained by the model, the rest are not covered. The results confirmed that there is no a simple linear model can be built to explain the relationship between these data.

The second experiment focused on utilization data information for classification. From the previous figure, we can see that different cities in USA have different economy environment. Therefore, their unemployment rate, GDP and personal income are different. So, the second analysis aims to answer the question that whether the data can be used for classification of geo location. There classification methods are selected as logistic regression (probability based), random forest (tree based) and multi-layer perceptron (neuron network based). All three kinds of economic data information are

used as input data and the associated state information is used as label. The dataset is shuffled first then 80% data points are selected as training set while the rest are used as testing set. If a state with less than 2 data points, the state will be filtered off. After training and testing, a series of results are obtained and summarized in the table below. From the results, we can conclude that random forest model provides the best performance. However, none of them give a result with good quality. The logistic regression is significantly worse than the others. Only about 20% of data can be classified reliably using GDP, personal income and unemployment rate information but it still indicates that geo-location is associated with the economic data at certain level. One of defects of the project design is that the state is used as label. In the final datasets used for training and testing, there are 38 states reserved as classes. Regarding to only a few thousand data points in the datasets, the classes are too many. I should include a better geo-location label such as divided US into 4 major areas as SE (south east), SW (south west), NE (north east), NW (north west) or 6 classes as NE, SE, NW, SW, MN (mid-north), MS (mid-south) which definitely will provide a better result. Also, I only used data from year 2008, 2009, 2014 and 2015 due to I run the program on local machine. If deploying this on a cluster with more computing resources, more data from different years can be included which will provide a more enhanced dataset for training respecting to how to improve the performance.

Model	Precision	Recall	f-score
Multi-layer perceptron	0.165	0.174	0.169
Logistic regression	7.15e-6	0.003	1.43e-5
Random forest	0.261	0.238	0.249

Conclusion

As a conclusion, this project aims to establish a relationship between geo location information and economic data. GDP, personal income and unemployment rate information are collected from different resources and the data are integrated to generate a data frame linked the economic data with the associated geo location information namely state and city. The integrated data are used for a simple classification analysis to investigate if the economic data is strong enough to classify the geo-location namely state. Three models are trained with best performance as 0.249 (F-score) from the random forest model. There are many improvement spaces exist to obtain better classification results. One of the most import contribution of this project is that all the process including data integration and classification are written in Scala with spark framework which allow flexible scaling to include significant large size of data. Through finishing this project, the map-reduce-aggregation process is used many times to integrate the data and simple classification methods are used which all covered by this data science course.