

Machine Learning Project  
  
Predicting Salary from Cost of living Index

By

Muhammad Tafani Rabbani

&

Hadi Nourallah

Introduction

In this project we would like to identify of the possibility of predicting an average of salary in a city by knowing the Cost of Living Index. Cost of Living Index is data that usually used by people who want to move to other city.

Cost of living index (COLI) provides data of prices in a city, the purpose of it to make easier of calculating expense living in a city. Make comparison of data, people can see through the difference of the price between city.

The 3 main data that is provided by COLI are rent index, groceries index, and restaurant index. The index is a percentage of price comparison with prices in New York, USA. For example if a city has groceries index of 30, it means the groceries price of that city is 30% of price in New York.

The data that we use is coming from Numbeo.com cost of living index early 2017 data. The cost of living index table is free and published publicly in the internet, and so it’s not a sensitive data and we have the right to use it for this project.

Numbeo.com is a website that provide data and information provider of prices in city around the world, the data that they are getting are from public, any visitor of the website can help them by providing update of data of their city.

Hypothesis

As the main objective of this project, we would like to find the best possible way to predict an average salary of a city. By knowing the prices of restaurant, groceries, rent, and purchase power of a people living in that city, we would like to find a formula and seeing through it if the prediction salary is possible.

In this way, we would like to make it easier for people to analysis the market condition of their city compare of others, also for young people to know more about their future career condition.

Previous work

The previous study was done by Archit khosla in competition on kaggle. Using data of set of job ads published in the UK.  
The detail of the project can be check in https://cseweb.ucsd.edu/~jmcauley/cse190/reports/sp15/012.pdf

The Dataset

In this project we are using WEKA for machine learning study, the data that we are gathering from Numbeo.com are COLI table in HTML file, Average salary of country rank in HTML.

Numbeo.com is website that provide all the data for this project. In this project we use data of COLI table, Average salary of country rank and put it in csv file that we put in WEKA.  
  
 First of all, we use a normal method of how getting the HTML data out of the website, using browser inspection (“in our case is Google chrome”) we just copy paste the important tag of the HTML to a text file.

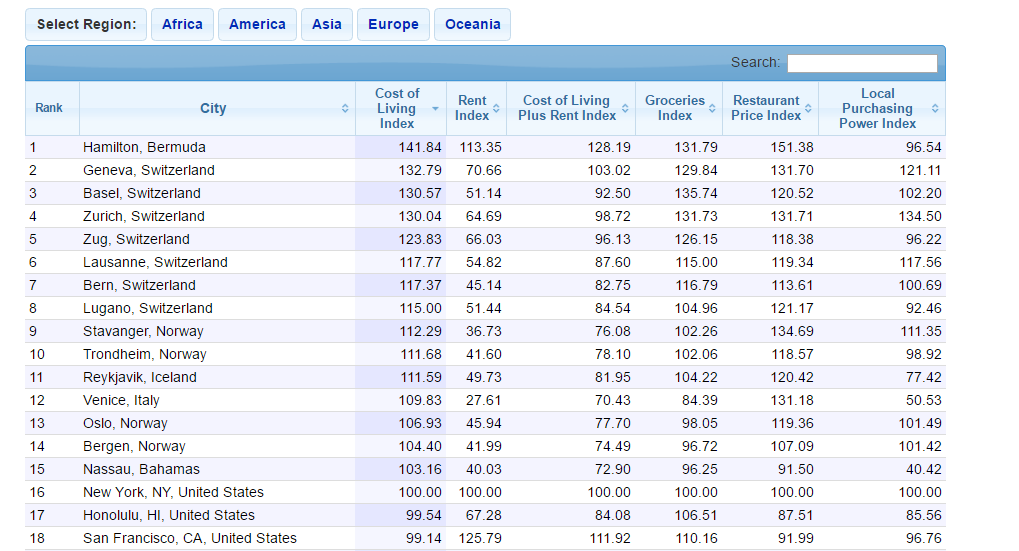
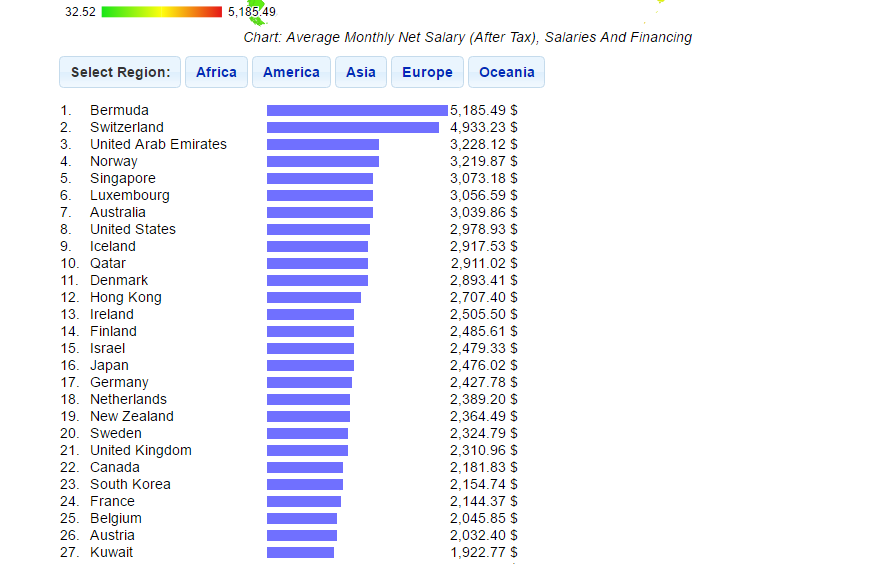
Second Step, we use text manipulation technique to change the HTML structure to JSON file and to CSV with the description of steps bellow:

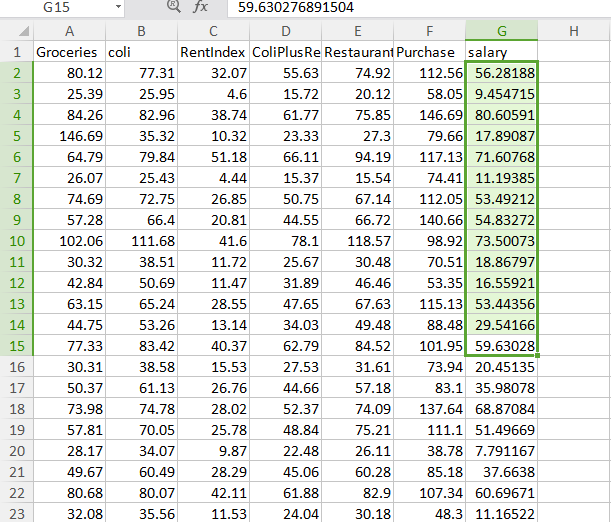
1. using tag, we read the tag and paste it to new file, in here since the HTML  
 file is ugly and same, we have to categorize which one is COLI, restaurant,   
 groceries, rent and purchase index using line pointer. We are using C++   
 programming for this step.  
  
 2. Changing the salary list of html file to JSON file, actually instead of taking   
 the html structure, by simply copy the text normally by labeling the text is faster. From there, again just changing the structure of the text using line   
 pointer. We are using C++ programming for this step.

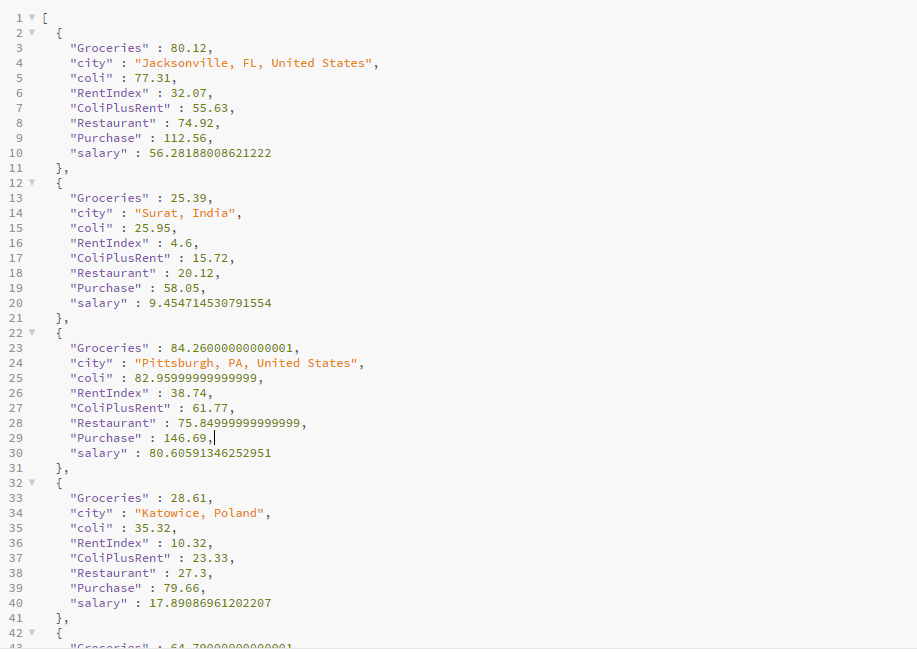
1. Combine the two JSON file, since in the first JSON is a city list, and second   
    JSON file is country list, we put the country information to the city of that country itself, we are using swift Programming for this step.
2. Changing the JSON file into CSV file. There is no programming involve in   
    this step, simply using online converter from JSON to CSV.

And that’s how we got the data that we will use it for our project in WEKA. Even tho WEKA has library of using JSON, we chose to make it simple by converting it to CSV file.

Cost of living index and salary list:



The final data JSON to CSV



Method of Machine Learning

Because of the Predicting Salary is our objective, we are using 2 methods of classification. Linear Regression and M5P tree method which is provided by WEKA.

Linear Regression Method

This method is by mean the simple one for predicting a numeric value, We are using 6 Attributes in this method, which are Groceries, coli, RentIndex, Restaurant, Purchase, salary. Using 20 folds of cross validation, Correlation coeficient of 0.9818, Mean absolute error 3.971, Root mean square root 5.302, the Relative absolute error of 16.44%, and Root relative square error of 18,9%.

This method as we mention is the simplest method, the result of it is a formula that is a linear regression of the data. Which WEKA found as

Salary =

0.192 \* Groceries +

0.5657 \* RentIndex +

0.1827 \* Restaurant +

0.3471 \* Purchase +

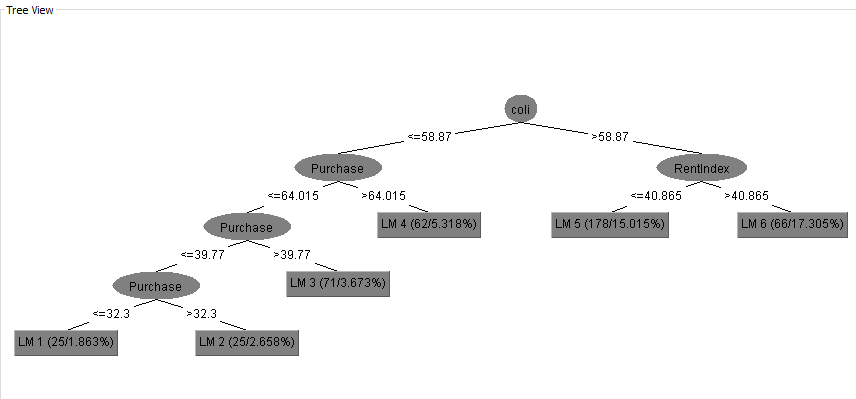
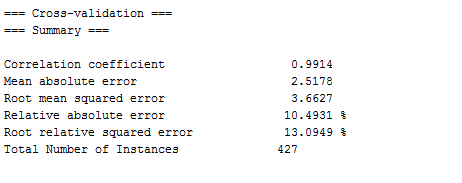
-25.0892

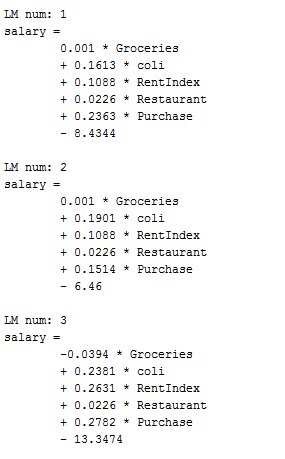
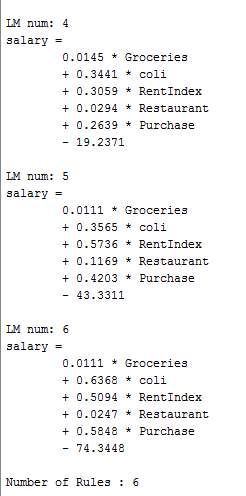
M5P tree Method

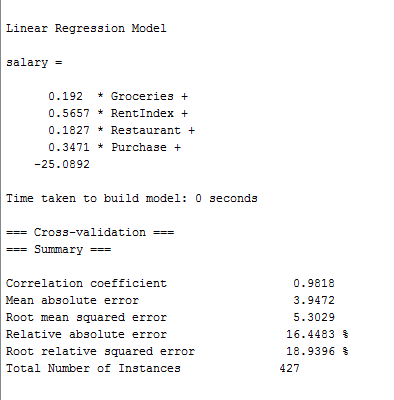
This is one of the extension of Linear Regression method, the difference here are to produce more than 1 formula and make a decision tree of which formula will be used. The attributes that is used is the same as the Linear Regression Method. The result is Correlation of coefficient 0.991, Mean Absolute error 2.552, Root Mean square error 3.72, Relative absolute error of 10.638% and Root Relative square error of 13.3109%.

Also this method produce 6 linear function within 6 rules.

M5P decision tree

M5P result



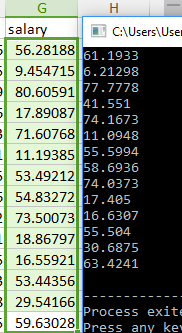
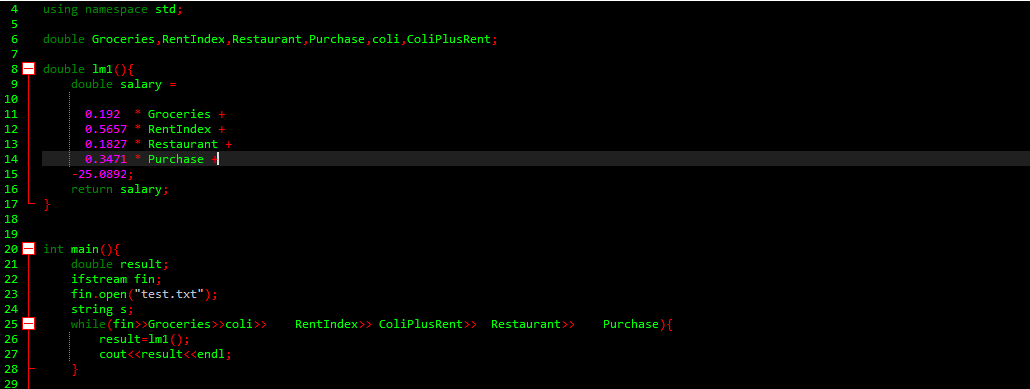
Linear Regression

Final Build

Once we did all the method, we will got the formula that we need for our application. Which mean the building process will not be really hard.  
  
Just as example, I will use c++ application to apply this formula. I made 2 application with linear regression and m5p tree method. I put some test sample for this example.

Linear Regression

This application is simply put a function that will using formula to calculate the input.

The result will be:

That is comparison of

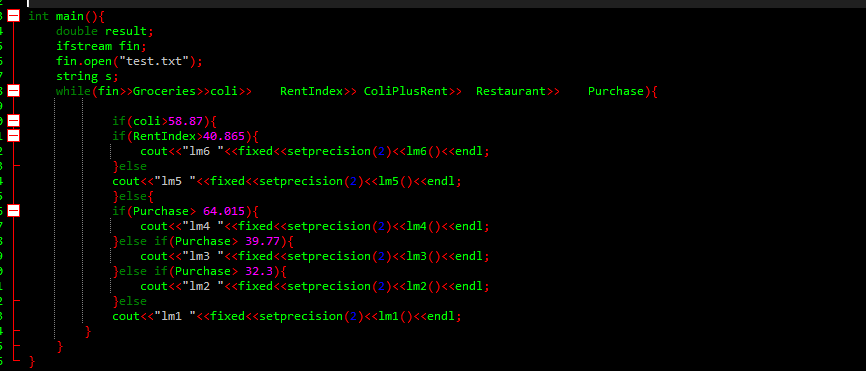
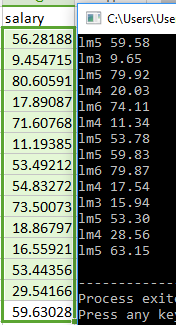
The sample (left) and

The result of formula

(right)

THE M5P application

Like previous example, this application will have 6 function that will calculate as the 6 formula that we got.

The result will be

As fellow picture:

Conclusion

Overall this approach is possible, eventho the result still has high percentage of errors, in the future it will has to developed with the population data, geographic location, even the profession of the people average salary will be helpful.