Analysis on 2013 American community using Regression model

Algorithms for Massive Datasets

Università degli Studi di Milano

Nallapaneni Aditya Sai (matr. 933570)

aditya.nallapaneni@studenti.unimi.it

Burka Kumar Sachin Kumar (matr. 934991)

sachinkumar.burkakumar@studenti.unimi.it

March 5, 2021

Abstract: The project illustrates the American community survey data of 1 to 50 states, with household records. We deal with only one variable which is HINCP which comes under housing unit and also we implement regression algorithm with ridge regression, Root mean square error and R2 function in reproductible and scalable form. Finally we conclude with exploratory data anaylsis through plotting actual and predicted values. All the material for this project is available on the following GitHub Entry

Keywords: Linear regression, Root mean square error, Regression , Ridge regression,large datasets

Contents

We declare that this material, which will be now submitted for assessment, is entirely our own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of our work. We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should we engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by us or any other person for assessment on this or any other course of study.

1 Data Description

The American Community Survey is an ongoing survey from the US Census Bureau. In this survey, approximately 3.5 million households per year are asked detailed questions about who they are and how they live. Many topics are covered, including ancestry, education, work, transportation, internet use, and residency.

There are two types of survey data provided ie housing and population and we considered housing dataset which each row is a housing unit, and the characteristics are properties like rented vs. owned, age of home, etc.

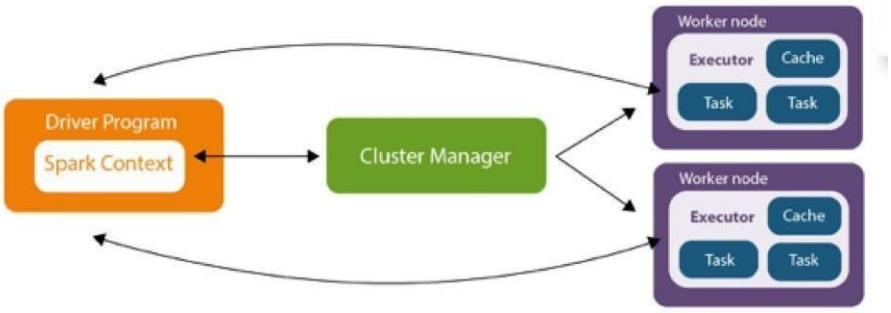
Each data set is divided in two pieces, ”a” and ”b” ,where ”a” contains states 1 to 25 and ”b” contains states 26 to 50. both the data sets have weights associated with them, Weights are included to account for the fact that individuals are not sampled with equal probably (people who have a greater chance of being sampled have a lower weight to reflect this).

There are total of 231 columns with ample entries in each housing datasets and the weights variable WGTP and Housing, Family income is represented as HINCP and FINCP.

2 Data Organisation

The Spark SQL DataFrames and Datasets APIs are used to process structured file data without the use of core RDD transformations and actions. This allows to analyze the structured data much faster than by applying the transformations on RDDs created. Spark is the linking component in a Unified Analytics framework, that encompasses data engineering (cleaning, transforming), data analytics (SQL), as well as machine learning and model deployment. Spark also allows SQL access to datasets that provides information about the structure of both the data and the computation that is being performed, and provides rich integration between SQL and regular Python, Java or Scala code, including the ability to join RDDs, represent a collection of items distributed across many compute nodes that can be manipulated in parallel, and SQL tables

Spark image:



Spark also allows SQL access to datasets, empowering analytics on so-called data lakes. Data pipelines in Spark can be tightly coupled with machine-learning algorithms; where Spark takes care of data management for both training and deployment of learned models. As such, Spark is the linking component in a Unified Analytics

framework, that encompasses data engineering (cleaning, transforming), data analytics (SQL), as well as machine learning and model deployment. when running SQL using another programming language, like it will be done on this project, the results will be returned in the form of a Dataset or DataFrame.

3 Data Preprocessing

In data preprocessing the main task is to clean and transform the data in the dataset, as of given in our dataset "2013 American Community Survey" here we have chosen two CSV files "housing a"(husa) and "housing b"(husb), Where "husa" carriers data of 1 to 25 states and "husb" 25 to 50 states, For the housing data, each row is a housing unit, and the characteristics are properties like rented ,owned, age of home, etc. Each of this files contains 231 columns with nulls values and husa contains 7.5 million entries and husb contains 7.2 million, as we can see its a large number of entries, we took 2 data frames for housing a and b. As mentioned in the data description out of 5 attributes given we consider only one as the predicted label and remaining 4 are removed, but we have 2 attributes in our housing dataset those are HINCP(Household income) and FINCP(family income) attributes, where all the values are of integers, So we remove one attribute by filtering the FINCP and HINCP missing values adjusted to value ADJINC, by using fillna() and the following missing values in CONP,SMP,MHP,MRGP,RNTP,SMOC,PINSP are adjusted to ADJHSG values and ACR,AGS,BROADBND,BUS,DIALUP,DSL,FIBEROP,VACS,MODEM,MRGI,MRGT,MRGX,OTHSVCEX,RNTM,VALP,SATELLITE,FES,FPARC,GRPIP,MV,NPF, OCPIP,SMX,TAXP,WKEXREL,WORKSTAT,FGRNTP,FSMOCP,WIF,GRNTP, ACCESS,BATH,BLD,FS,HANDHELD,HFL,LAPTOP,REFR,RMSP,RWAT, RWATPR,SINK,STOV,TEL,TEN,TOIL,VACS,VEH,YBL,FFINCP,FHINCP,HHL, HHT,HUPAC,HUPAOC,KIT,LNGI,MULTG,PLM,RESMODE,COMPOTHX, HUPARC,FULP,WATP,GASP,ELEP,NR,NRC these are adjusted to constant value '0' in both the data frames and the remaining attributes which are HUGCL,NOC,NPP, PARTNER,PSF,R18,R60,R65,SRNT,SSMC,SVAL,BDSP, are adjusted to the distinct values.

For the columns ELEP,GASP,FULP,WATP the values which are >= 3 are adjusted to the column 'ADJHSG' values, here we considered 4 data frames each attribute in each data frame for housing a and housing b. In total we merge all the data frames of husa and husb by using unionALL. So, df final is our final data frame with 1476313 entries containing housing a and housing b datasets as shown in fig.

```
[ ] df_final.printSchema()

root
 |-- RT: string (nullable = true)
 |-- SERIALNO: integer (nullable = true)
 |-- DIVISION: integer (nullable = true)
 |-- PUMA: integer (nullable = true)
 |-- REGION: integer (nullable = true)
 |-- ST: integer (nullable = true)
 |-- ADJHSG: integer (nullable = true)
 |-- ADJINC: integer (nullable = true)
 |-- WGTP: integer (nullable = true)
 |-- NP: integer (nullable = true)
 |-- TYPE: integer (nullable = true)
 |-- ACCESS: integer (nullable = true)
 |-- ACR: integer (nullable = true)
 |-- AGS: integer (nullable = true)
 |-- BATH: integer (nullable = true)
 |-- BDSP: integer (nullable = true)
 |-- BLD: integer (nullable = true)
 |-- BROADBND: integer (nullable = true)
 |-- BUS: integer (nullable = true)
 |-- COMPOTHX: integer (nullable = true)
```

4 Implementation of Regression Algorithm

4. Linear regression

Ordinary least squares Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the *x*-axis and *y*-axis should be linearly correlated.

Linear regression is used to predict a quantitative response Y from the predictor variable *X*.

Mathematically, we can write a linear regression equation as:

$$y = a + bx \tag{1}$$

Where *a* and *b* given by the formulas:

$$b(slope) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} \tag{2}$$

$$a(intercept) = \frac{n\sum y - b(\sum x)}{n} \tag{3}$$

Here, *x* and *y* are two variables on the regression line.

b = Slope of the line.

a = y-intercept of the line.

x = Independent variable from dataset

y = Dependent variable from dataset

LinearRegression fits a linear model with coefficients $w = (w1, \ldots, wp)$ to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation.

We use Ridge regression(Ridge regression addresses some of the problems of Ordinary Least Squares by imposing a penalty on the size of the coefficients with l2 regularisation).

4.1 Ridge regression

Ridge regression is a special case of Tikhonov regularization in which all parameters are regularized equally it is a technique for analyzing multiple regression data that suffer from multicollinearity. Where multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value.

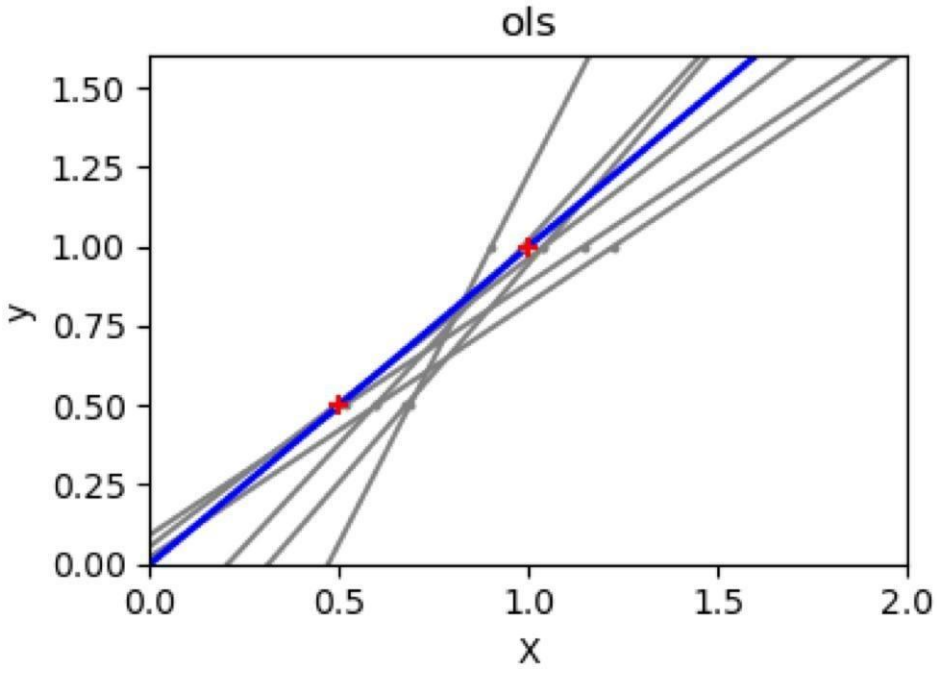
Ridge regression addresses some of the problems of Ordinary Least Squares by imposing a penalty on the size of the coefficients. The ridge coefficients minimize a penalized residual sum of squares:

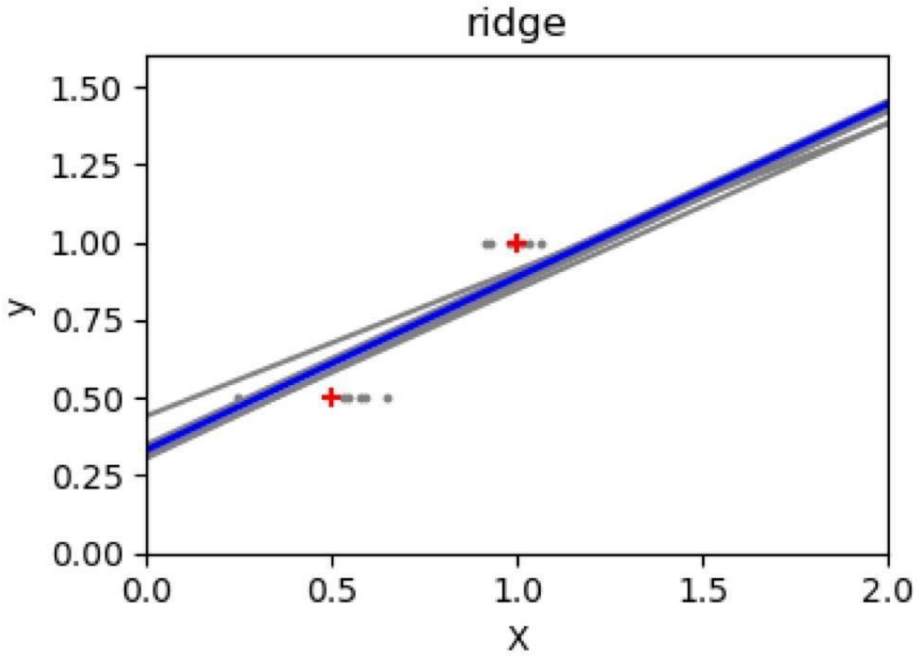
$$\min_{w} || Xw - y ||_2^2 + \alpha || w ||_2^2 \tag{4}$$

This model solves a regression model where the loss function is the linear least squares function and regularisation which is given by the l2-norm. This estimator has built-in support for multi-variate regression (i.e., when *y* is a 2d-array of shape (n_samples, n_targets)).

When lambda > 0 (i.e. regParam > 0) and alpha = 0 (i.e. NetParam = 0) ,then the penalty is an L2 penalty Which is condition for Ridge regression.

Ridge regression is basically minimizing a penalised version of the least-squared function. The penalising shrinks the value of the regression coefficients. Despite the few data points in each dimension, the slope of the prediction is much more stable and the variance in the line itself is greatly reduced, in comparison to that of the standard linear regression


ols
1.50
1.25
1.00
0.75
0.50
0.25
0.00
y
0.0
0.5
1.0
1.5
2.0
X


ridge
1.50
1.25
1.00
0.75
0.50
0.25
0.00
y
0.0
0.5
1.0
1.5
2.0
X

4.2 RMSE (Root mean squared error)

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are, RMSE is a measure of how spread out these residuals, In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in regression analysis to verify experimental results. This is the same as MSE (Mean Squared Error) but the root of the value is considered while determining the accuracy of the model.

Hence,

$$MSE = \frac{1}{N}\sum_{i=1}^{n} (\text{actual values} - \text{predicted values})^2 \quad (5)$$

Here *N* is the total number of observations/rows in the dataset. The sigma symbol denotes that the difference between actual and predicted values taken on every *i* value ranging from 1 to *n*.

The errors are squared before they are averaged. This basically implies that RMSE assigns a higher weight to larger errors. This indicates that RMSE is much more useful when large errors are present and they drastically affect the model's performance. It avoids taking the absolute value of the error and this trait is useful in many mathematical calculations. In this metrics lower the value, better is the performance of the model.

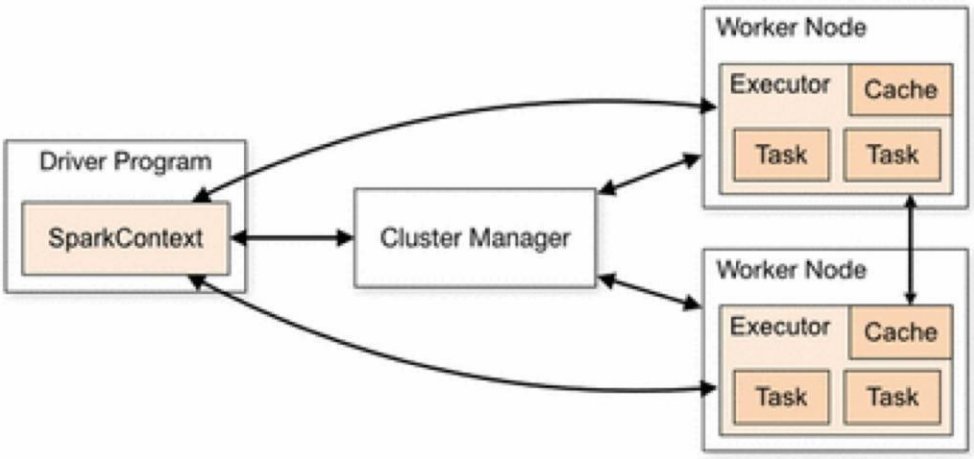
4.3 R^2

It is also known as the coefficient of determination. This metric gives an indication of how good a model fits a given dataset. It indicates how close the regression line (i.e the predicted values plotted) is to the actual data values. The R squared value lies between 0 and 1 where 0 indicates that this model doesn't fit the given data and 1 indicates that the model fits perfectly to the dataset provided.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2} \quad (6)$$

5 Scalabiliy

The large amounts of data have created a need for new frameworks for processing. The MapReduce model is a framework for processing and generating large-scale datasets with parallel and distributed algorithms. Apache Spark is a fast and general engine for large-scale data processing based on the MapReduce model. The main feature of Spark is the in-memory computation. Spark is optimized for speed and computational efficiency by storing most of the data in memory, it can underperform Hadoop MapReduce when the size of the data becomes so large that insufficient RAM becomes an issue.



Linear processing of huge datasets is the advantage of Hadoop MapReduce, while Spark delivers fast performance, iterative processing, real-time analytics, graph processing, machine learning and more..A task is the smallest unit of work that Spark sends to an executor. SparkContext represents a connection to a computing cluster.

6 Description of experiments

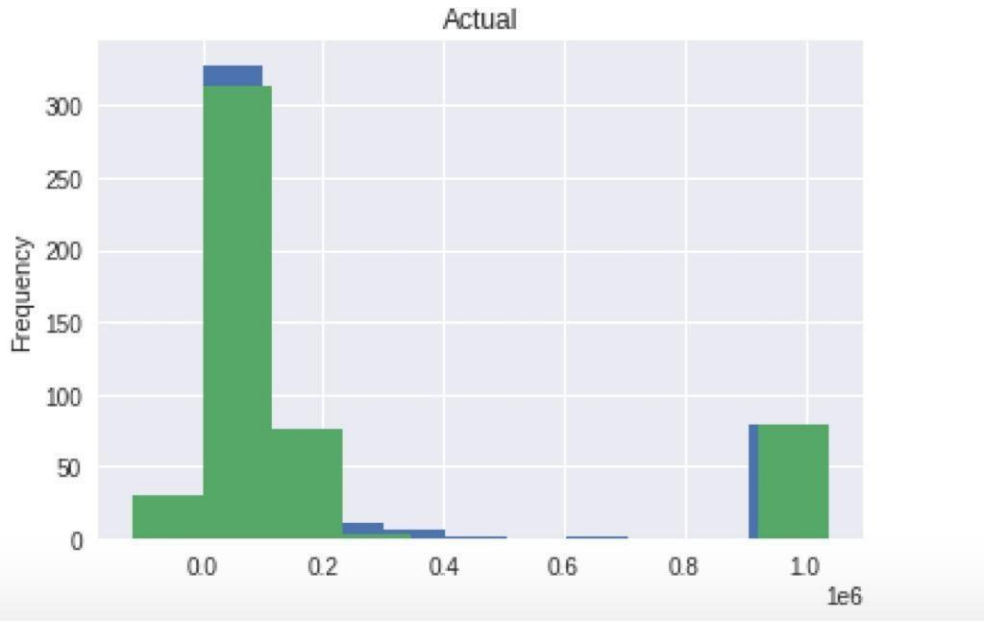
The dataset is splitted into random with 80% as train set and 20% as test sets. Our input features for the linear regression task includes 231 columns while the label is the HINCP column which refers to the Housing Income across 50 states in USA. As we know HINCP is the label given which is independent, we train the linear regression model using ridge regression.

Ridge regression is a simple technique to reduce model complexity and prevent overfitting which may result from simple linear regression and the cost function is altered by adding a penalty equivalent to square of the magnitude of the coefficients, the penalty term (lambda) regularises the coefficients such that if the coefficients take large values the optimisation function is penalised. So, ridge regression shrinks the coefficients and it helps to reduce the model complexity and multi-collinearity, when $\lambda \to 0$, the cost function becomes similar to the linear regression cost function, So lower the constraint (low λ) on the features, the model will resemble linear regression model. But in our case while training the model using metric RMSE ie Root Mean Square Error which is the general validation metrics and the r2 which is Performance evaluation metric (The R squared value lies between 0 and 1 where 0 indicates that this model doesn't fit the given data and 1 indicates that the model fits perfectly to the dataset provided) which predicts the value which are close to the actual values, but In our case we have got high RMSE value, but given the data available we have some outliers .Therefore, RMSE is not a desirable evaluation metric.

As we can see the standard deviation value is very high, either we can drop them from results or we can keep them. Dropping them will get the RMSE value to minimum but with the cost of information loss. If we keep them we have to accept the cost for high RMSE with a tradeoff of scalability and precision.

7 Results and Discussion

We have done a comparison of actual and predicted label, where mean is almost same and standard deviation is less than the actual which implies to more precise our model is and we also perform exploratory data analysis to view the labeled and predicted data which shown below:



Here Blue refers for actual values and green for predicted one, So the model is better performed with some outliers.So, as we know that we have only one independent variable we can absolutely use linear regression model for better prediction and results, Moreover we performed RMSE and R2 metrics but we see R2 performs better in our dataset.

References

[1] https://www.kaggle.com/census/2013-american-community-survey

[2] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

[3] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html#sklearn.linear_model.Ridge

[4] https://spark.apache.org/docs/latest/ml-classification-regression.html#linear-regression

[5] https://runawayhorse001.github.io/LearningApacheSpark/reg.html#ridge-regression

[6] https://scikit-learn.org/stable/modules/linear_model.html#ridge-regression

[7] https://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html#sphx-glr-auto-examples-linear-model-plot-ols-py