Prediction of Obesity Risk Level Among Youth

Ketki Vyas, Ritika Joshi, Sanchita Gawand

Department of Information Systems, California State University

Los Angeles

Tel. 323-343-2916, Fax. 323-343--5209

e-mail : [kvyas2@calstatela.edu](mailto:kvyas2@calstatela.edu), [rjoshi5@calstatela.edu](mailto:rjoshi5@calstatela.edu), [sgawand@calstatela.edu](mailto:sgawand@calstatela.edu)

**Abstract:** The paper explains the method and process used for prediction of obesity risk level among youth. The major concentration of the project is to create a prediction model with Microsoft azure, Databricks and Oracle on youth behavior surveillance system for 'Obesity Overweight and Weight Control' using specific features to determine the prevalence of health risk behaviors. Prediction model is created using multiple algorithms in all three environments, further Root Mean Square and Accuracy is compered.

**1. Introduction**

This project uses Microsoft Azure, Databricks and oracle to create prediction model for obesity risk and uses multiple algorithms in reach environment to compare and get the closest predicted value. The data set of YRBSS (Youth Risk Behaviors Surveillance System) monitors priority health risk behaviors that contribute markedly to the leading causes of death, disability, and social problems among youth and adults in the United States. As per their study, there are six categories of priority health behaviors among youth and young adults that contribute to unintentional health problems, injuries and violence: 1. Tobacco use 2. Alcohol and other drug use 3. Sexual behaviors ​4. Obesity, Overweight and Weight Control  ​5. Physical inactivity​ 6. Dietary Behaviors​

We have data related to all the six categories listed above among which we have selected to create prediction model for data related to “Obesity, Overweight and Weight Control”. The reason why we selected Obesity related data for our project because the prevalence of obesity was 39.8% and affected about 93.3 million of US adults in 2015~2016. The estimated annual medical cost of obesity in the United States was $147 billion in 2008 US dollars. Also, the medical cost for people who have obesity was $1,429 higher than those of normal weight. In this model we are Predicting column “Greater Risk Data Value” for Obesity, Overweight and Weight Control data. Higher the number in the column more the risk of obesity.

**2. Related Work**

We came across a project by NCBI (National Center for Biotechnology Information) Predicting childhood overweight and obesity using maternal and early life risk factors. There is one more model by the NCBI Dynamic Model Predicting Overweight, Obesity, and Extreme Obesity Prevalence. In this case Mechanistic insights can be provided from a mathematical model. The objective of this study is to model known multiple population parameters associated with changes in body mass index (BMI) classes and to establish conditions under which obesity prevalence will plateau. Trends Several prediction models exist, but most have not been externally validated or compared with existing models to improve predictive performance.

**3. Specifications**

The dataset comprises of data collected by YRBSS (Youth Risk Behaviors Surveillance System) which monitors priority health risk behaviors that contribute markedly to the leading causes of death, disability, and social problems among youth and adults in the United States. The dataset is of the size 1.6 GB and covers data from 1991-2017. Table 1 shows files and size of the files from dataset.

Table 1 Data Specification

|  |  |
| --- | --- |
| Size | 1.6 GB |
| URL | <https://www.kaggle.com/raylo168/dash-yrbss-hs-2017> |
| File Type | CSV |

Table 2 Azure Specification

|  |  |
| --- | --- |
| Storage | 10 GB |
| Node | 1 |
| Python Version | 2.7.11 |
| IP address | [https://studio.azureml.net](https://studio.azureml.net/) |

Table 3 Databricks Specification

|  |  |
| --- | --- |
| Storage | Cluster 5.2 (includes Apache Spark 2.4.0, Scala 2.11) |
| Memory | 6GB Memory,  0.88 Cores, 1 DBU |
| Node | 3.5.2 |
| Python Version | 2.7.11 |
| IP address | <https://community.clo>[ud.databricks.com/](https://community.cloud.databricks.com/) |

Table 4 Oracle Specification

|  |  |
| --- | --- |
| Storage | 682GB, 12 OCPUs |
| Memory | 180 GB |
| Node | 6 |
| Python Version | 2.7.14 |
| IP address | [129.150.127.176](mailto:yourusername@129.150.127.176) |

**4. Predictive Modelling**

The spark Machine Learning has been used in Databricks to find out the model accuracy with cross validation and to find the root mean Square error. Three different algorithms have been used to compare the model’s accuracy and to find the best fit model. The same model has also been built for prediction in Microsoft Azure and Oracle. The following sections provides the model and the result.

**4.1 Microsoft Azure Machine Learning Studio**

Here we have used three algorithms naming Decision Forest Regression, Linear Regression and Boosted decision Tree Regression to build the model and predict the outcome. The sample data file has been created from youth\_risk\_behaviour.csv file as Azure has limit on file size. Next select columns in dataset model was used to select the columns which will be used in prediction. Later data cleaning was done, and an algorithm was applied to verify the model to get root mean square.



Figure 1. Microsoft Azure Model



Figure 2. Microsoft Azure Model continued

In order to predict the level of obesity risk as precise as possible, 'Filter Based Feature Selection' is chosen. Targeted eight best columns that had effect on 'Greater Risk Data Value'. Greater Risk Data value is the labelled data.



Figure 3. Feature based selection

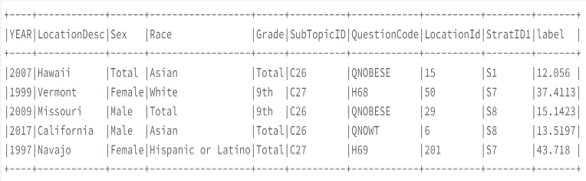


Figure 4. Features

Algorithm Used: Decision Forest Regression

Root Mean Square Error: 3.90

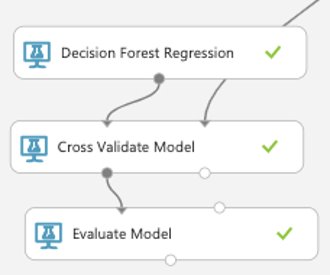


Figure 5. Decision Forest Regression Model

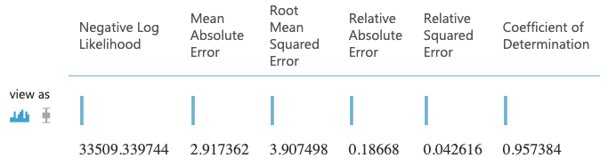


Figure 6. Decision Forest Regression Result

Algorithm Used: Linear Regression

Root Mean Square Error: 7.66

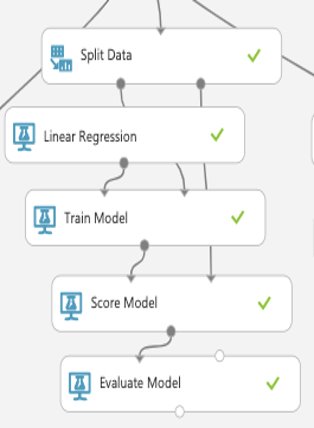


Figure 7. Linear Regression Model

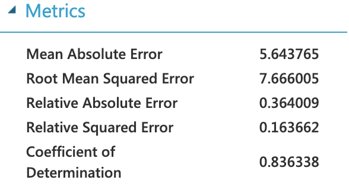


Figure 8. Linear Regression Result

Algorithm Used: Boosted Decision Tree Regression

Root Mean Square Error: 3.84

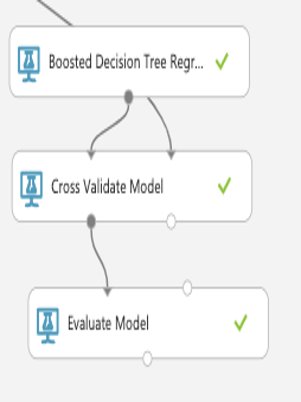


Figure 9. Boosted Decision Tree Regression Model

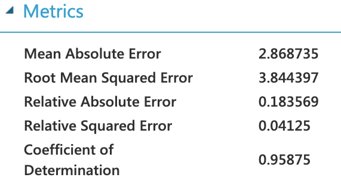


Figure 10. Boosted Decision Tree Regression Result

**4.2 DATA Bricks (SPARK Machine Learning)**

Algorithm Used: Linear Regression Model

Accuracy: 11.02

Root Mean Square Error: 66.05



Figure 11. Linear Regression Code

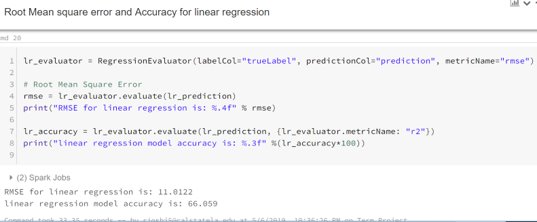


Figure 12. Linear Regression RMSE and Accuracy

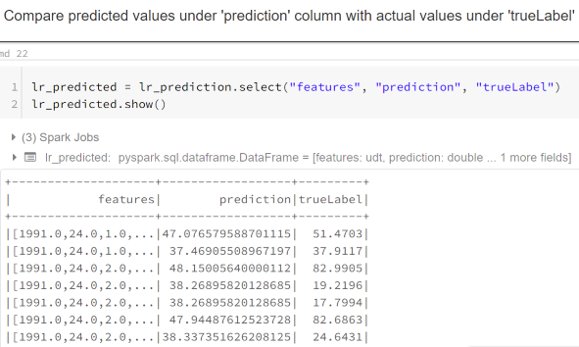


Figure 13. Linear Regression Comparison

Algorithm Used: Decision Tree Regression Model

Accuracy: 93.80

Root Mean Square Error: 4.705



Figure 14. Decision Tree Regression Code

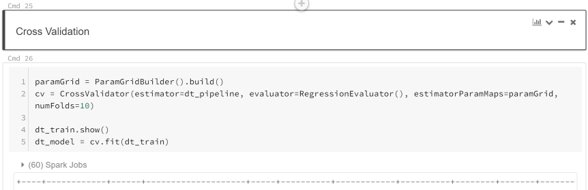
****

Figure 15. Decision Tree Regression Cross validation

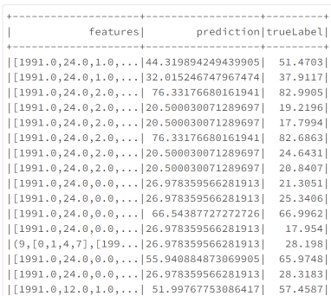
****

Figure 16. Decision Tree Regression Comparison

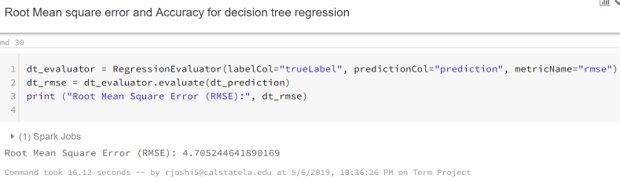
****

Figure 17. Decision Tree Regression RMSE

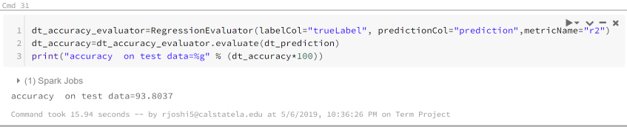
****

Figure 18. Decision Tree Regression Accuracy

**4.3 Oracle (SPARK Machine Learning)**

Algorithm Used: Linear Regression Model

Accuracy: 11.02

Root Mean Square Error: 66.05

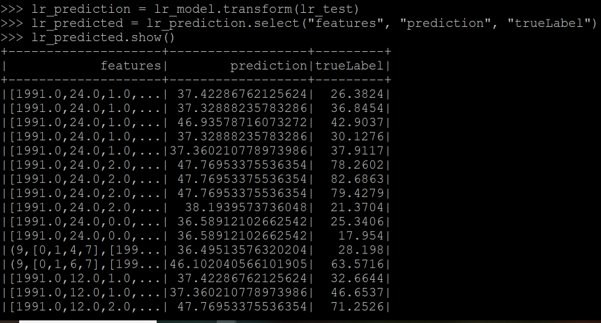


Figure 19. Linear Regression Predicted Value Comparison

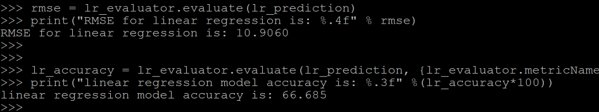


Figure 20. Linear Regression Accuracy and RMSE

Algorithm Used: Decision Tree Regression Model

Accuracy: 94.27

Root Mean Square Error: 4.521

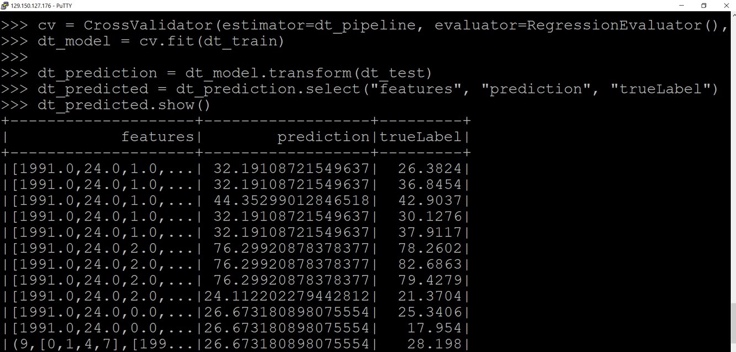


Figure 21. Decision Tree Predicted Value Comparison

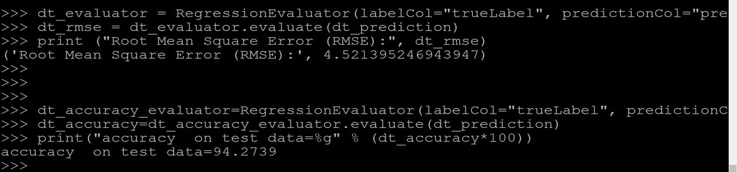


Figure 22. Decision Tree Accuracy and RMSE

**5. Summary**

We have successfully used the tools learnt in class such as data bricks and Microsoft Azure to build predictive models and to analyses and visualize the result. We have created a summery table (Table 5) to compare Linear Regression and Decision Forest Tree regression across all three environments, in our case we found that Decision Tree Regression is best fit and gives the best result. For more information on Prediction Model and code visit project’s GitHub link[[1]](#footnote-2).

Table 5 Databricks Specification

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **​Platform​** | **​Dataset​**  **Size​** | **Satistics​** | **Linear ​**  **Regression​​** | **DecisionTree**  **Regression​​** | **ExecutionTime​** |
| Azure​​  ​​ | 90 MB​ | RMSE​​ | ​7.66​ | ​3.90​  ​ | 4:45 min​ |
| Databricks​​ | 1.6 GB​ | RMSE​​ | ​11.07​ | ​4.502​ | 15:22 min​ |
| ​ | Accuracy​​ | ​66.06​ | ​94.403​ | ​ |
| Oracle​​ | 1.6 GB​ | RMSE​​ | ​10.90​ | ​4.502​ | 23 min​ |
| ​ | Accuracy​​ | ​66.68​ | ​94.27​ | ​ |

**6. References**

### [1] Overview | Yrbss | Adolescent and School Health | Cdc

<https://www.cdc.gov/healthyyouth/data/yrbs/overview.htm>

[2] Classification and Regression - Rdd-based Api

<https://spark.apache.org/docs/2.1.0/mllib-classification> regression.html

[3] Building A Linear Regression with Pyspark and Mllib

Susan Li-Susan Li

<https://towardsdatascience.com/building-a-linear-regression-with-pyspark-and-mllib-d065c3ba246a>

[4] Machine Learning with Spark and Python George Radescu

<https://blog.epigno.systems/2018/02/18/machine->learning-with-pyspark-linear-regression/

1. GitHub Link: <https://github.com/rjoshi5/Obesity-Risk-Level-Among-Youth> [↑](#footnote-ref-2)