Predictive Analysis of salary For Different Job Titles

Ashwin Karthik (306598726) – akarthi@calstatela.edu
Shanmathi ArulMurugan (306594267) – sarulmu@calstatela.edu
Kaushik Sridharan (306611492) – ksridha@calstatela.edu
Department of Information Systems, California State University Los Angeles
Mentor**: Dr. Jongwook Woo**

Abstract: The analysis is on the data about the employee salary information in the New York city which contains features like Salary, Job ID, Hours/Shift, Agency, etc. Based on these features we have predicted the salaries of people who would work in the specific places in future. The dataset has more than 5 million records. We have used a few regression and classification models for predicting the lables and checking its accuracy for better performance.

1. Introduction

Dataset we used for this project:

- DOB-Job-Application-Filings

Dataset is in CSV format and size is about 2.7GB.
This dataset contains employee salary information of New York city in United States. Sample features in the dataset are Job ID, Hours/Shift, Agency, Business title, Part-time and Full-time Indicator, Salary etc. The aim of our project is to predict the Income of an employee based on the available features from the dataset by utilizing Machine Learning Algorithms and build accurate models using Azure ML, Spark ML.
The Machine Learning Algorithms we used are:

- Boosted Decision Tree Regression
- Decision forest Regression
- Linear Regression

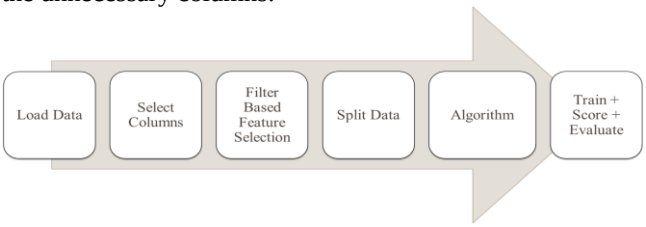
2. Services

Services used in this project are Azure ML, Databrics community edition, Spark ML, Python and Tableau. We used Apache spark Version – Spark 2.3.0(Scala 2.11), which has Memory 6 GB(0.88 cores, 1DBU). We also used Data Bricks File System.

3. Work Flow

3.1 AzureML

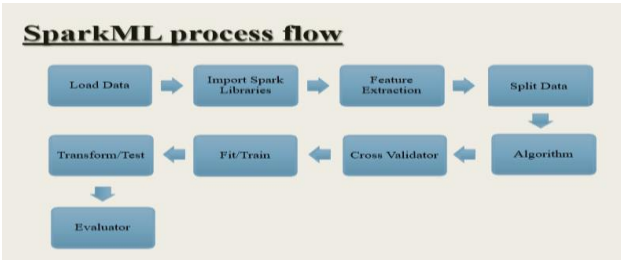
To start with, we first loaded the data to Azure ML, then we selected the columns, which is also a part of data cleaning as we took only columns which we are going to use further. Then we used Filter Based Feature Selection to filter out the unnecessary columns.



Thereafter to split the data into training set and testing set we used Split Data module. As mentioned earlier we have used different algorithm to proceed further. Lastly, we Train, Score, and Evaluate the model to find the best model.

3.2 SparkML

The first step in Workflow of SparkML is to load the data. Then we have imported Spark Libraries, followed by extracting features from raw data, Removing Outliers and splitting data. Post that we did Cross validation for training and transforming so that testing can be performed on it. Last step includes Evaluation in the workflow of SparkML.



4. Background Work

We have used datasets to analyze the features that act as major factors affecting the salary of the employee's state wide. Based on these factors we have come up with model which would help in predicting the Salary of the employees working at New York. Based on these factors we develop a regression model to predict Total Salary.

5. Analysis and Visualization

5.1 Feature selection Analysis

We have used the filter-based feature selection analysis on Azure ML using the Spearman Correlation method.

Spearman Correlation compares the non-linear features with the labels and produces a score between the range -1 to 1. Values which are around -1 and 1 have a strong

correlation whereas, if the range is around 0, it has a weak correlation and it can be omitted. That is how we found out the best features that can be used.

5.2 Analysis of the data: AzureML

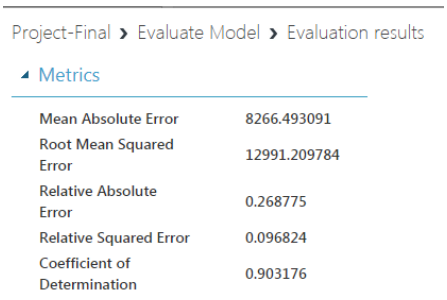
Boosted Decision Tree Regression: Boosted Decision Tree Regression module is used to create an ensemble of regression trees using boosting. Boosting means that each tree is dependent on prior trees, and learns by fitting the residual of the trees that preceded it. Thus, boosting in a decision tree ensemble tends to improve accuracy with some small risk of less coverage. We have used Salary RangeTo, Hours/Shift, Agency Posting type, Business title, Title code, Level, Job Category, Full-time/Part-time indicator, Salary frequency and Work Location. Result we got for RMSE is 12436.852182 and COD is 0.914188

Metrics

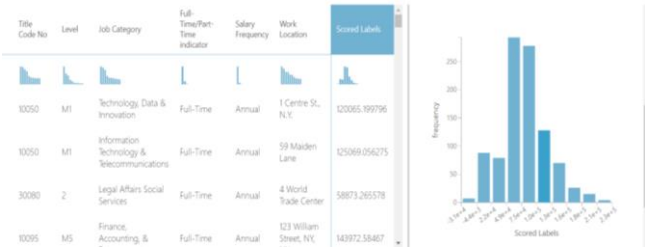
Mean Absolute Error	7675.41378
Root Mean Squared Error	12436.852182
Relative Absolute Error	0.243189
Relative Squared Error	0.085812
Coefficient of Determination	0.914188

Boosted Decision Tree Regression using cross validator:

We have used the same process as for Boosted decision tree regression but with cross validator to check whether it provides more better results. But we have found that the results of the other methods have better results than this. So we have considered this as an optional method.The results are as shown in the image below.



Linear Regression: In simple **linear regression** a single independent variable is used to predict the value of a dependent variable. In multiple **linear regression** two or more independent variables are used to predict the value of a dependent variable. The difference between the two is the number of independent variables. We have used Salary RangeTo, Hours/Shift, Agency Posting type, Business title, Title code, Level, Job Category, Full-time/Part-time indicator, Salary frequency and Work Location. The result we got for RMSE is 9961.10068 and COD is 0.944952



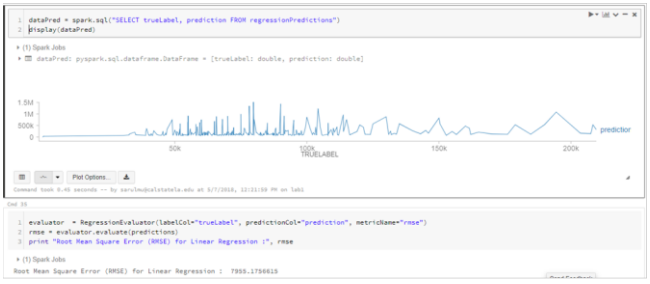
Metrics

Mean Absolute Error	4394.189922
Root Mean Squared Error	9961.10068
Relative Absolute Error	0.139226
Relative Squared Error	0.055048
Coefficient of Determination	0.944952

5.5 Spark ML Model

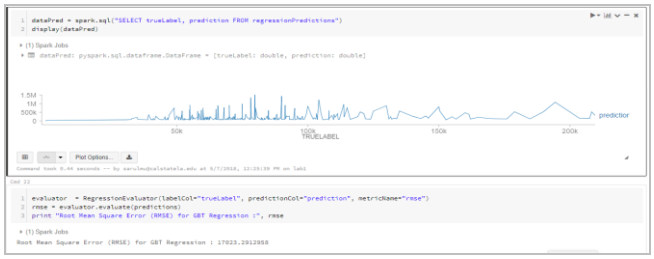
Apache Spark is a fast and general-purpose cluster computing system. It provides high-level APIs in Java, Scala and Python, and an optimized engine that supports general execution graphs. We have used Linear Regression and GBT Regressor algorithms. Features used in the modelling are Job_ID, Posting_Type, Title_Code_No, Level, Hours_Shift, FullTime_PartTime, Salary_Frequency.

Linear Regression: The Root Mean Square Error (RMSE) of New York city incomes using Linear Regression Algorithm is 7955.1756615.



GBT Regressor RMSE: The R Mean Square Error (RMSE) of New York city incomes using GBT Regressor Algorithm is 17023.2912958.

.



6. Summary and Conclusion

Based on selected features predictive analytics was conducted to predict the Total Salary of employees working in New York City

Comparison between models in Azure ML and Spark ML is done and it is found that Spark's Linear Regression Provides the best predictive values.

Azure ML		Spark ML	
Boosted Decision Tree Regression	Linear Regression	Gradient Boosted Tree Regression	Linear Regression
RMSE=12436.852182	RMSE=9961.10068	RMSE=17023.2912958	RMSE=7955.1756615

7. Limitations & Challenges Faced

- Feature prediction was much easier in AzureML than that of SparkML
- If the data set has more details such as work experience , skill set etc., we could have predicted more accurate salary for different job titles, because the employee may be a fresher or an experienced candidate.

Dataset URL

https://data.cityofnewyork.us/Housing-Development/DOB-Job-Application-Filings/ic3t-wcy2

GitHub URL

https://github.com/koushiksri1994/CIS5560

References

- https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/
- https://people.apache.org/~pwendell/spark-nightly/spark-master-docs/latest/ml-classification-regression.html
- https://github.com/apache/spark/tree/master/examples/src/main/python/mllib