Machine Learning with LA County Restaurant Violations

Vignesh Srinivas, Natya Srinivasan, Abhishek Shah
Department of Information Systems, California State University
Los Angeles
vravish@calstatela.edu
nsriniv@calstatela.edu
ashah36@calstatela.edu

Abstract: Our dataset contains "Los Angeles County restaurants/markets violations". It has the inspection data for violations of restaurants and markets for the years 2014,2015 and 2016. It contains various fields with information such as Inspection date, Name of Market/Restaurant, City, Violation Code & Violation description, Final Score, Grade (A/B/C/SC) for the restaurant/market. We planned to build a supervised learning model to predict the critical violations that will affect the overall grade/score of a Market/Restaurant, where we trained and learned the models and made the Predictive analysis using the same. We have used **Regression** as well as **Classification** algorithms, and predicted which model has better accuracy.

1. Introduction

More than 54 billion meals are served in 884,400 restaurants in the USA every year. 46% of the American money spent on food is for restaurant meals. On an average, 44% adults eat at a restaurant on any typical day of a calender in the United States. Of a mean 550 **foodborne disease** outbreaks are reported to the Disease Control and Prevention Center. Thus, our task is to prevent restaurant-associated foodborne diseases by helping Restaurants in avoiding Violations.

The goal of this project is to create a platform to analyze and visualize the grades/scores of the restaurants in LA county by the violations they have made each day using machine learning models. This predictive analysis helps the Restaurants to easily identify the score/grade they would get for their mistakes, i.e. violations they have made, resulting into improvisation in those violations as soon as possible.

2. Related Work

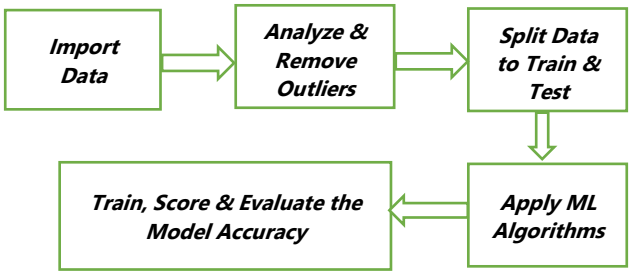
A paper on foodsafetynews.com mentions about how the local public health department inspects the restaurants in all the counties in different states. They talk about how frequent inspections take place and what things (violations) the inspectors look for. However, it fails to show the penalty points for violations made by a restaurant which our dataset sets an edge. An article in the LA Times about Health Inspections which mentions about the conversations between co-workers talking about mismanagement in the department. They say that they were directed to do more inspections, because the numbers were more important to management than quality work. They ended up with no good results. On the other hand, our analysis is better in terms of qualitative as well as quantitative aspects as we mention everything with specifics.

3. Background/Existing Work

We have used the existing dynamic dataset of LA County restaurants to carry out detailed predictive analysis on restaurant violations according to the inspections made. Basically, the idea is to bring awareness to the restaurants about the reviews they get and the things they need to improve by limiting the violations they made in the past.

4. Our Work

We intend to make a predictive analysis on grading/scoring the restaurants based on the violations they have made using Azure ML and Spark ML. We built a supervised learning model using Regression and Classification techniques. Below is the work flow of our project.



The **cluster details of Spark ML** are as follows:

- Apache Spark Version, Spark 2.1 (Auto-Updating, Scala 2.10)
- Memory – 6GB, 0.88 cores, 1 DBU (Data Brick unit)
- File System – DBFS (Data Bricks File System)

A. Preprocessing Steps (Analyzing the Data)-

ACTIVITY DATE	ACTIVITY MONTH	ACTIVITY DATE	ACTIVITY YEAR	NAME	SITE CITY	VIOLATION CODE	POINTS	SCORE	GRADE
9/2/2015	9	2	2015	#3 TOMMY'S BURGER	DUARTE	16F027	1	91	A
9/2/2015	9	2	2015	#3 TOMMY'S BURGER	DUARTE	16F040	1	91	A
9/2/2015	9	2	2015	#3 TOMMY'S BURGER	DUARTE	16F029	1	91	A
9/2/2015	9	2	2015	#3 TOMMY'S BURGER	DUARTE	16F014	2	91	A
9/2/2015	9	2	2015	#3 TOMMY'S BURGER	DUARTE	16F044	1	91	A
9/2/2015	9	2	2015	#3 TOMMY'S BURGER	DUARTE	16F037	1	91	A
9/2/2015	9	2	2015	#3 TOMMY'S BURGER	DUARTE	16F030	1	91	A
9/2/2015	9	2	2015	#3 TOMMY'S BURGER	DUARTE	16F033	1	91	A

We analyzed a part of our data and found out that a Restaurant named Tommy's Burgers which was inspected on 09/02/2015 has made 8 violations and got a total score of 91 and their grade is A. It's total penalty points is 9. Each Violation has a penalty point of either 0, 1, 2 or 4. The grade of a restaurant is calculated as below,

Grade Chart

Grade	Score Range
A	90 - 100
B	80 – 89
C	70 – 79
SC	< 69

B. Preprocessing Steps (Transforming the Data) -

In this preprocessing step, we transform the data according to our needs. We fired a query to group activity date with the name of the restaurant to find the number of violations made by that restaurant on a particular day of inspection. After grouping, our data looked like below,

Project - Regression › Apply SQL Transformation › Results dataset

rows 140470 columns 6

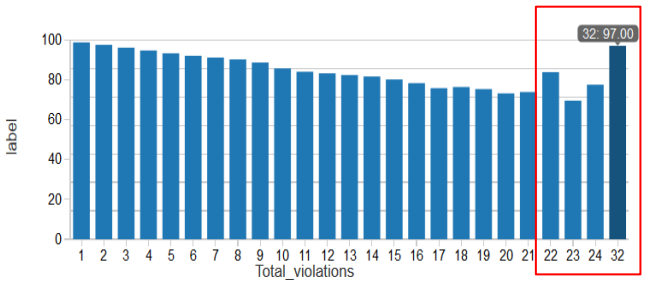
ACTIVITY DATE	NAME	No_of_Violations	GRADE	SCORE	Violation_points
2015-03-10T00:00:00	#1 BUFFET	5	A	92	8
2015-10-02T00:00:00	#1 BUFFET	7	A	93	7
2016-02-09T00:00:00	#1 BUFFET	6	A	90	10
2016-05-24T00:00:00	#1 BUFFET	6	A	90	10
2014-10-15T00:00:00	#1 CAFE	10	B	85	15
2015-05-27T00:00:00	#1 CAFE	5	A	94	6
2015-10-28T00:00:00	#1 CAFE	7	A	92	8
2016-05-25T00:00:00	#1 CAFE	5	A	92	8
2016-08-18T00:00:00	#1 CAFE	8	A	90	10

Group fields to find No_of_Violations

For example, restaurant #1 BUFFET made 5 violations on 3/2015, scored 92 and got A grade. While restaurant #1 CAFÉ made 10 violations on 10/2014, scored 85 and got B grade.

C. Preprocessing Steps (Analyzing the Trend and Removing the Outliers) –

Comparing the Average Score of all restaurants with the number of violations they have made, we came to a result that if the Score of a restaurant is high, then the number of Violations they have made is less. However, we see some inconsistencies in our data for total violations more than 21. These are the **Outliers** in our data**.**



Outliers

For example, if we have a look at the last bar, we see that 32 violations yields a score of 97, which is incorrect. Thus, we eliminated these Outliers (total violations more than 21) using a SQL query.

D. Machine Learning: Regression –

The goal of Regression model is to predict the Score of a restaurant based on the Total Violations they have made. Its features are Total Violation counts for a restaurant on a day of inspection and the label is Score, i.e. the Final score of a restaurant.

Here the Algorithms we used are:

Azure ML – Linear Regression, Boosted Decision Tree Regression.

Spark ML – Linear Regression, Decision Tree Regression.

Regression using Azure ML

Project - Regression › Evaluate Model › Evaluation results

Metrics

Mean Absolute Error	1.155625
Root Mean Squared Error	1.608544
Relative Absolute Error	0.409593
Relative Squared Error	0.19472
Coefficient of Determination	0.80528

Metrics

Mean Absolute Error	1.188872
Root Mean Squared Error	1.635237
Relative Absolute Error	0.421377
Relative Squared Error	0.201236
Coefficient of Determination	0.798764

Boosted Decision Tree Regression (Accurate)
Vs
Linear Regression

Here, the Evaluation metric is RMSE (Root Mean Squared Error). Split of 70% for Train and 30% for Test. From the analysis, we found out Boosted Decision Tree Regression to be better and Accurate model since its RMSE is **1.608** which is lesser than the RMSE of Linear Regression which is **1.635**.

Regression using Spark ML

```
evaluator1 = RegressionEvaluator(labelCol="trueLabel", predictionCol="prediction", metricName="rmse")
rmse1 = evaluator1.evaluate(prediction1)
print "Root Mean Square Error (RMSE) For Linear Regression Model:", rmse1
```

(1) Spark Jobs

Root Mean Square Error (RMSE) For Linear Regression Model: 1.63698905742

Command took 1.93 seconds -- by vravish@calstatela.edu at 5/4/2017, 7:53:28 PM on TEST

Linear Regression Model

```
evaluator2 = RegressionEvaluator(labelCol="trueLabel", predictionCol="prediction", metricName="rmse")
rmse = evaluator2.evaluate(prediction2)
print "Root Mean Square Error (RMSE) For Decision Tree Regression Model %g" % rmse
```

(1) Spark Jobs

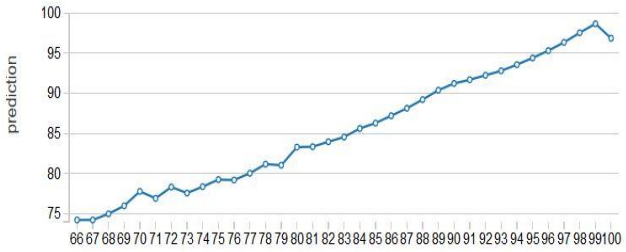
Root Mean Square Error (RMSE) For Decision Tree Regression Model 1.60965

Command took 2.03 seconds -- by vravish@calstatela.edu at 5/4/2017, 7:57:48 PM on TEST

Decision Tree Regression Model (Accurate)

In Spark, our Evaluation metric is RMSE (Root Mean Squared Error). Split of 70% for Train and 30% for Test. On comparing the Linear Regression with Decision Tree Regression, we came up with a conclusion that Decision Tree Regression is a more Accurate model since its RMSE is **1.609** which is less than the RMSE of Linear Regression which is **1.636.** Below is the Line chart of True Label vs

Prediction for Decision tree model. We see a constant increase in the line, which depicts a perfect regression line.



True Label Vs Prediction

E. Machine Learning: Classification –

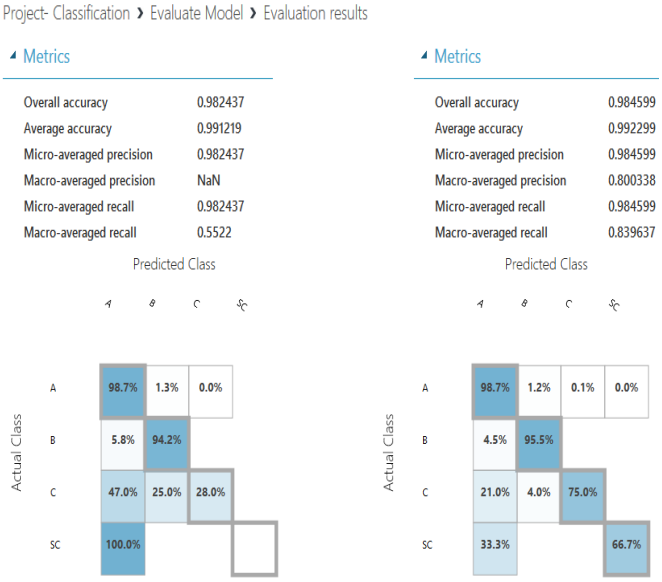
The goal of Classification model is to predict the Grade of a Restaurant based on Total Violations and their Penalties. Its features are Total Violations & Penalty points and the label is Grade, i.e. the Final Grade of a restaurant.

The Algorithms used are:

Azure ML – Multi-class Decision Jungle, Multi-class Decision Forest.

Spark ML – Decision Tree Classifier, Random Forest Classifier.

Classification using Azure ML



Multiclass Decision Jungle
Vs
Multiclass Decision Forest (Accurate)

Evaluation metrics used here are the Average Accuracy and the Confusion Matrix. Split of 70% for Train and 30% for Test. Out of the 39000 total restaurant inspections, 37500 got an A grade, 2000 got B, less than 100 in C & SC grade combined. So eventhough the predictions of C & SC are incorrect in model1, it got accuracy 99.1%. However, on comparing the Macro-Average recall value; the Multi-class Decision jungle had only **55% recall** value whereas the Multi-class Decision Forest has **83% recall** value. Therefore, its concluded that Multi-class Decision Forest is an accurate model.

Classification using Spark ML

```
evaluator = MulticlassClassificationEvaluator(labelCol="trueLabel", \
 predictionCol="prediction", metricName="accuracy")
accuracy = evaluator.evaluate(predictions)
print "Average Accuracy =", accuracy
print "Test Error = ", (1 - accuracy)
```

```
(2) Spark Jobs
Average Accuracy = 0.997646779178
Test Error =  0.00235322082244
```

Random Forest Classification model (Accurate)

```
evaluator2 = MulticlassClassificationEvaluator(labelCol="trueLabel", \
predictionCol="prediction", metricName="accuracy")
accuracy2 = evaluator2.evaluate(predictions2)
print "Average Accuracy =", accuracy2
print "Test Error = ", (1 - accuracy2)
```

```
(2) Spark Jobs
Average Accuracy = 0.996893673527
Test Error =  0.00310632647254
```

Decision Tree Classification model

Evaluation Metric in Spark that we used is Average Accuracy. Split of 70% for Train and 30% Test. From the executed analysis, we can make out that the Random Forest Classification model is better in Accuracy which is **99.76%** than Decision Tree classification model which is of accuracy **99.68%**.

F. Overview of Accurate Models –

	Azure ML	**Spark ML**
Regression	Boosted Decision Tree *RMSE:1.608*	Decision Tree *RMSE:1.609*
Classification	Multiclass Decision Forest *Accuracy:99.2%*	RandomForest Classification *Accuracy:99.7%*

Table showing the results of the accurate models

This table represents the results of the evaluation metrics of Accurate models in Regression and Classification models with Azure Machine Learning and Spark Machine Learning.

5. Observations

We have observed certain facts during our implementation with Azure ML and Spark ML. Feature prediction and Data preprocessing was easy in Azure ML on comparison with Spark ML. Wide range of Multi-class classification algorithms are available in Azure ML, whereas Spark ML consists of only 2 algorithms which we have used in our experiment. Evaluation Metrics like Confusion matrix, calculating precision, recall is not available for Multi-class classification in Spark ML. Also, Data visualization is easier in Spark ML.

6. Conclusion

The LA County Restaurant Grade/Score predictions using Regression and Classification are compared in both Azure ML and Spark ML (Databricks). On comparison, we found out that the predictions were almost similar in both Azure ML and Spark ML.This also helps in identifying the various accurate models for our dataset.

7. References

Articles & Papers – Health inspections

[1] http://articles.latimes.com/keyword/health-inspections

[2] http://www.foodsafetynews.com/restaurant-inspections-in-your-area/#.WRkeUWjytPY

Other References

[3] https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-evaluate-model-performance

[4] http://stackoverflow.com/questions/33636944/preserve-index-string-correspondence-spark-string-indexer

Dataset URL

https://data.lacounty.gov/Public-Health/LOS-ANGELES-COUNTY-RESTAURANTS-AND-MARKETS-VIOLATI/b9ey-v6ni

GitHub URL

https://github.com/arshah137/CIS5560-ML