Yelp Restaurant Review Analysis with Big Data Technologies

Nushrat Jahan Ria
nria1@lsu.edu
Louisiana State University, Baton Rouge, LA

Project Description

- It focuses on the analysis of Yelp's Business Dataset for predicting business success by taking into consideration customer reviews, star ratings and review counts, together with location data applying big data technologies.

- Data is processed in a distributed way using PySpark. PySpark offers efficient processing and transformation at scale through the DataFrame API that makes working with structured data easy.

- The train and evaluate some machine learning models, such as Logistic Regression, Stochastic Gradient Descent, Gradient Boosting, and Random Forest on business performance prediction. Thus, the integration of Dask within this scope will be affirmative to efficient parallel computing and memory optimization for tasks that need scaling beyond local memory constraints, complementing PySpark's processing.

- Cassandra is employed for the distributed storage system, whereas Kafka is used to allow real-time data streaming, thus laying the bedrock for real-time analytics of continuously updating Yelp reviews.

Development Details

- The project therefore begins by ingesting and cleaning the data; hence, the Yelp business dataset shall be loaded into PySpark DataFrames through missing value handling, renaming columns, and preparing the data for analysis.

- PySpark is used to conduct EDA that explains, in particular, the main features regarding star rating distribution and the counts of reviews to bring out trends and insight into the performance at businesses.

- It designs a machine learning pipeline that trains the models, such as Logistic Regression, SGD, Gradient Boosting, and Random Forest, on the data for the prediction of business success based on customer feedback.

- The performance of the models is estimated by accuracy, precision, recall, and F1-score evaluation metrics, where the Random Forest and Gradient Boosting ensemble methods perform better.

- Big data frameworks are integrated into this system, including PySpark, Dask, Cassandra, and Kafka, where data could be scalablely processed, efficiently stored, or even streamed in real time for future development.

Big Data Frameworks

- PySpark: This is designed for large-scale distributed data processing, fast in-memory computation, and parallel processing across clusters; hence, it is indispensable in efficiently processing large-scale datasets.

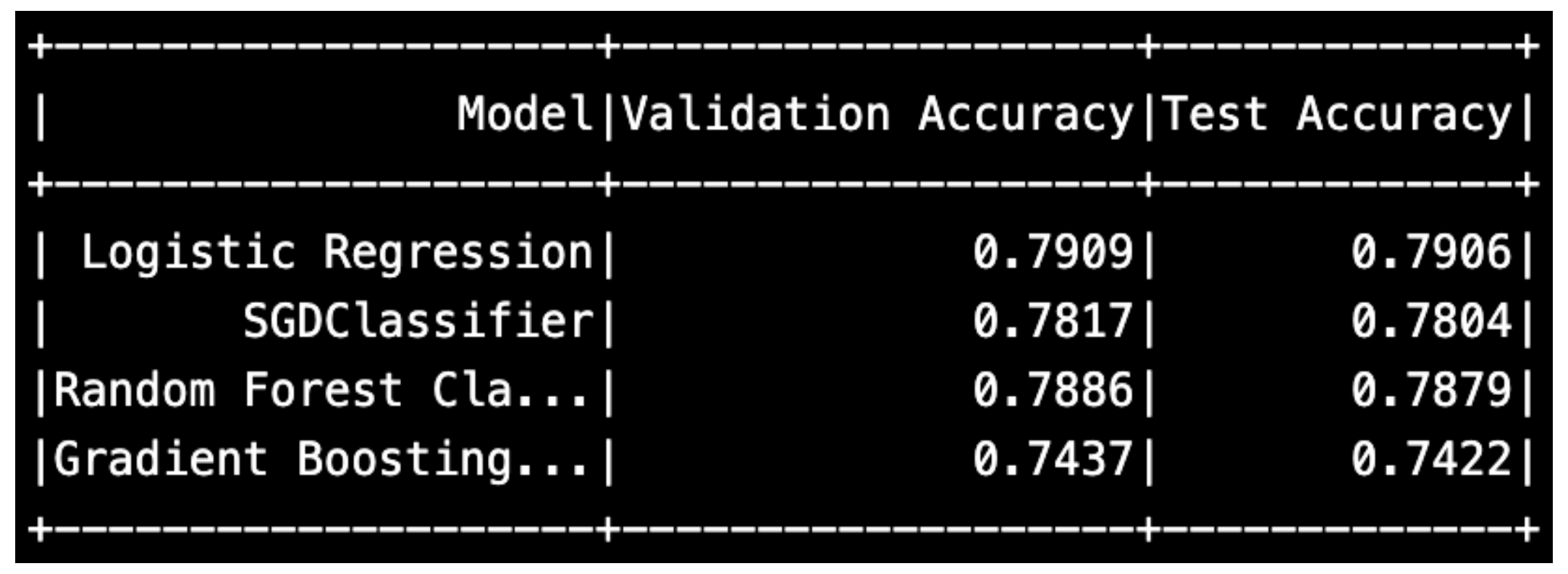
- Dask: This is for parallel computation when tasks do not fit into memory. Integrated with PySpark, it allows for scalable data analysis through efficient handling of heavy computation.

- Cassandra: This is a distributed NoSQL database intended to provide scalable data storage, featuring high availability and fault tolerance. The architecture is fitting for this project since it very efficiently stores and retrieves large volumes of structured business data.

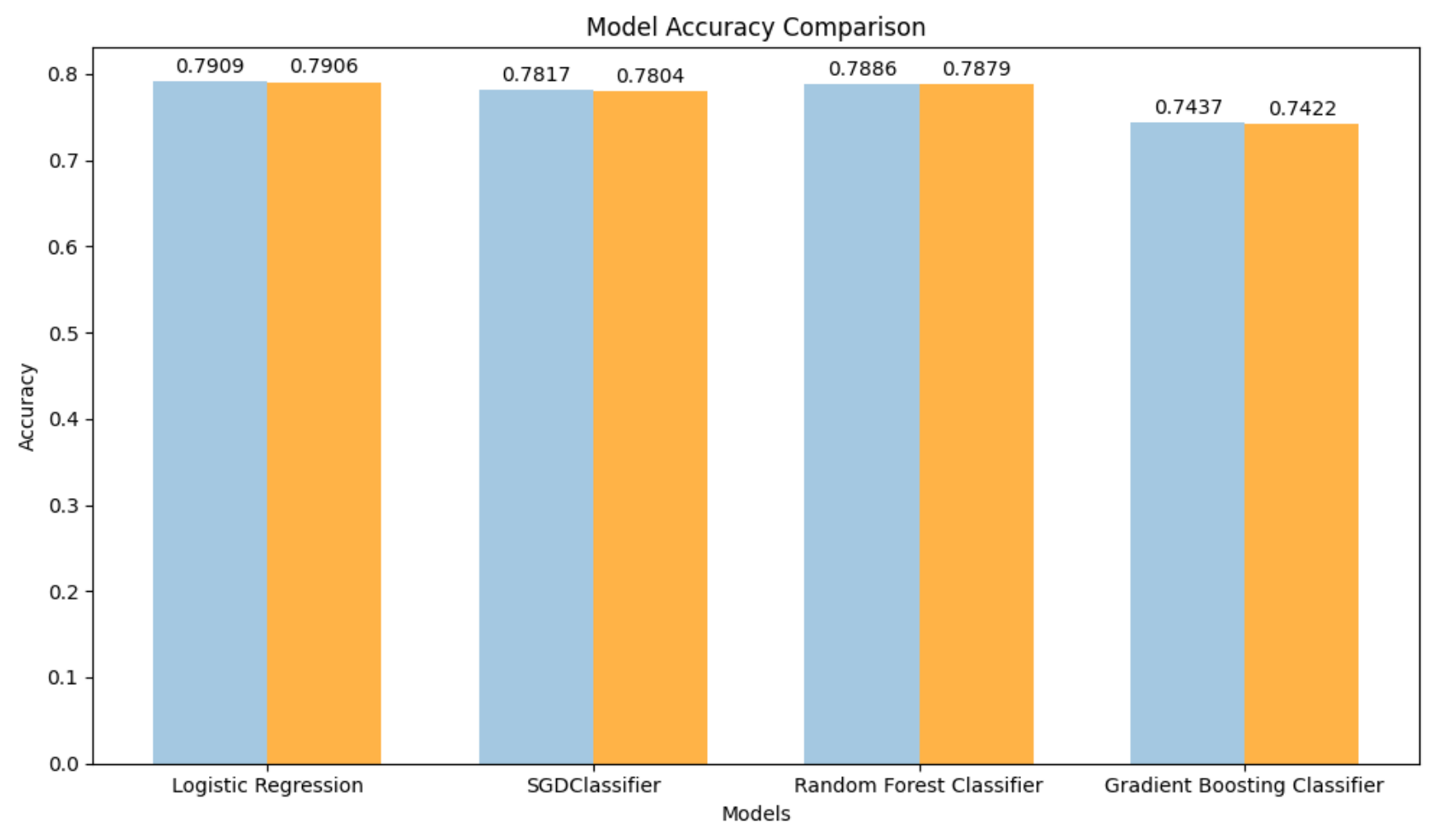
- Kafka: This becomes useful for real-time data streaming. In fact, Kafka allows for the continuous ingestion and processing of streams of data, thus laying the bedrock for any future real-time Yelp review analysis.

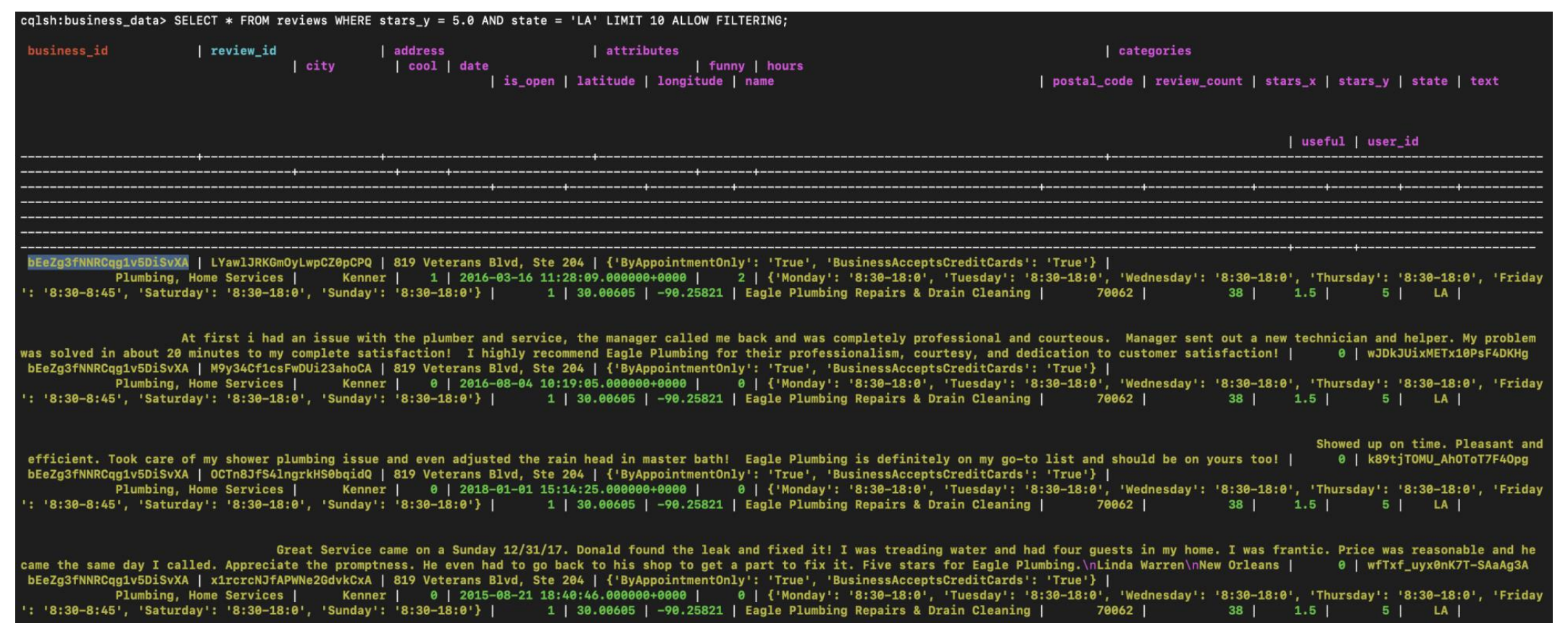
- Cassandra Driver: Cassandra is a Python library that interfaces with Cassandra and allows for efficient access and modification of data within the distributed storage system. This especially plays an important role when working with big data sets, such as those provided by Yelp.

Main Results & Screenshots

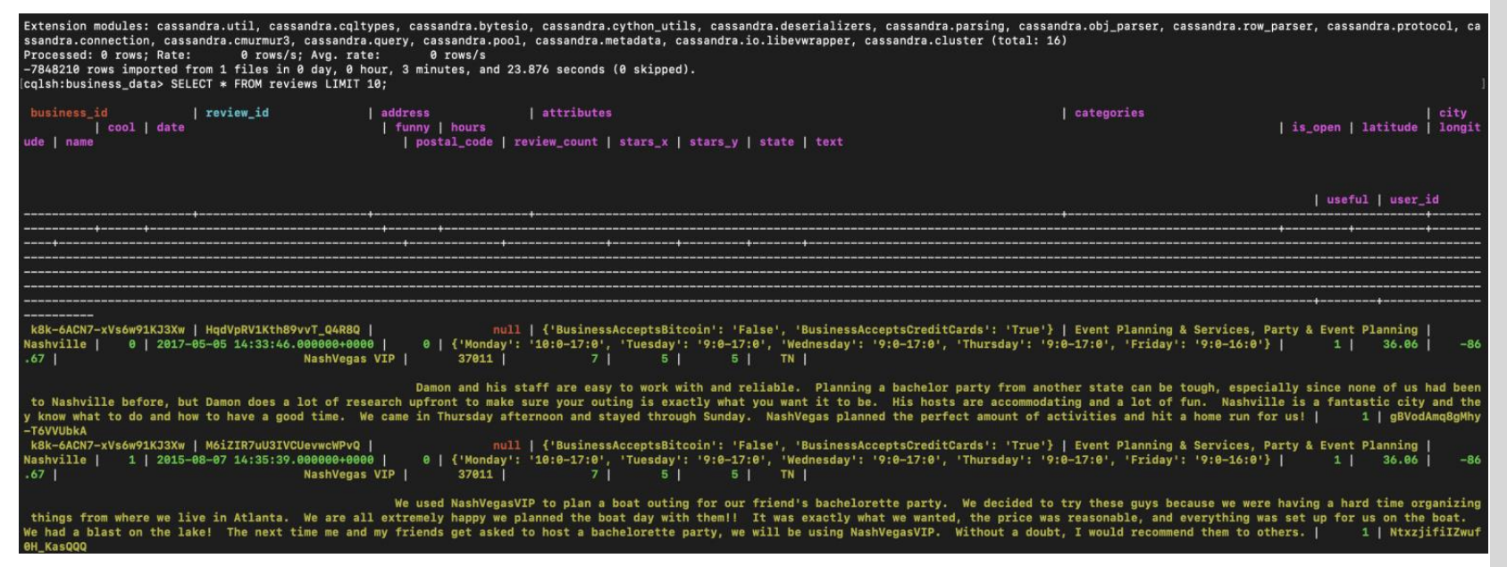


```
+--------------------+-------------------+-------------+
|               Model|Validation Accuracy|Test Accuracy|
+--------------------+-------------------+-------------+
| Logistic Regression|             0.7909|       0.7906|
|       SGDClassifier|             0.7817|       0.7804|
|Random Forest Cla...|             0.7886|       0.7879|
|Gradient Boosting...|             0.7437|       0.7422|
+--------------------+-------------------+-------------+
```

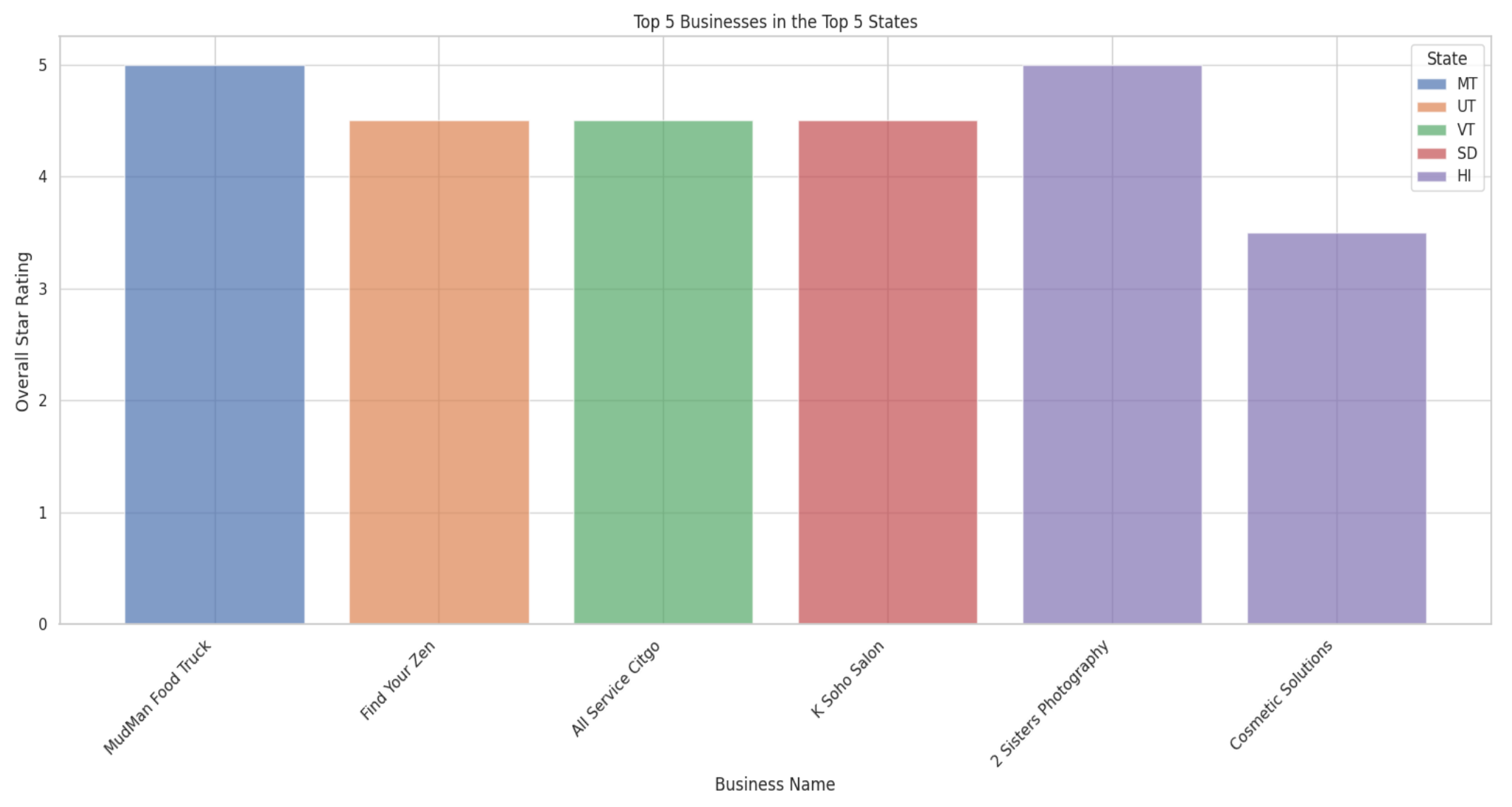


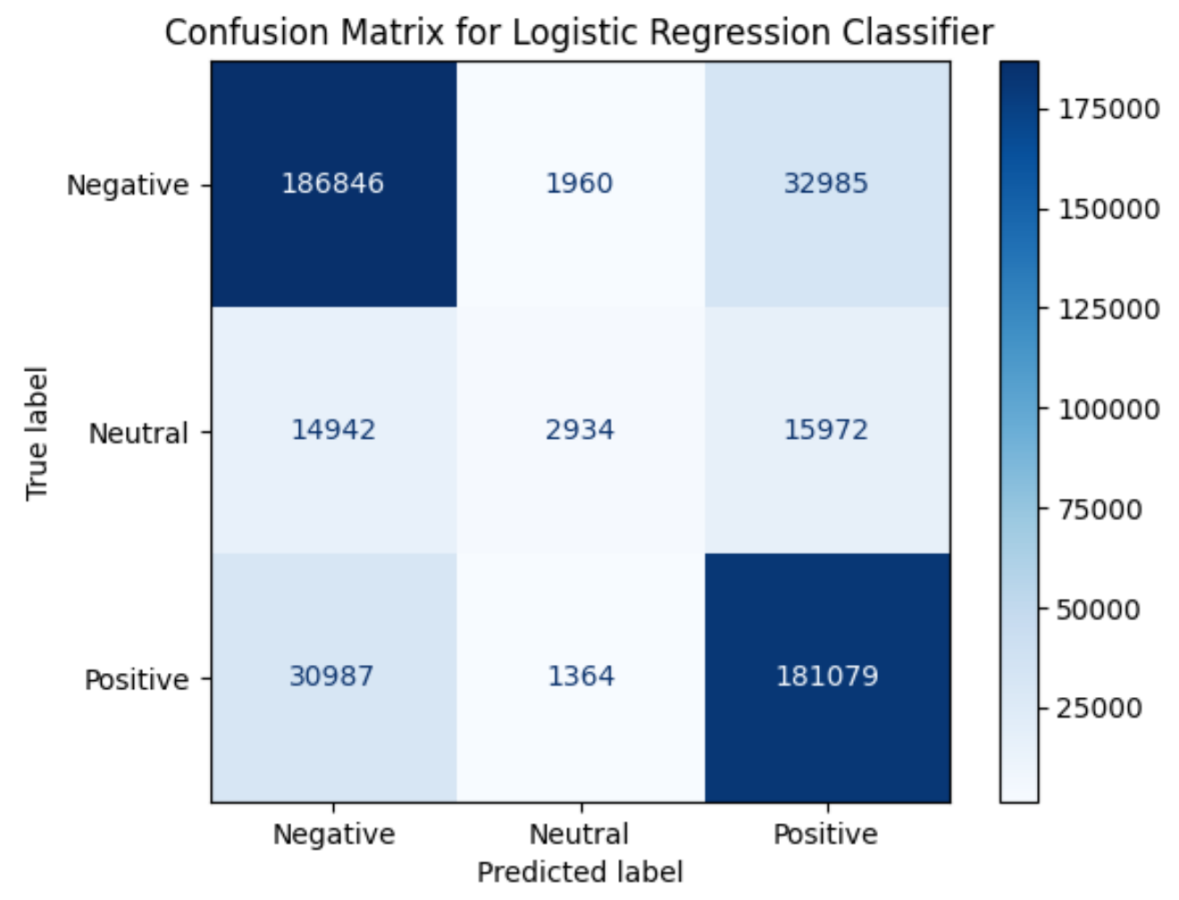


```
cqlsh:business_data> SELECT * FROM reviews WHERE stars_y = 5.0 AND state = 'LA' LIMIT 10 ALLOW FILTERING;

 business_id            | review_id              | address                  | attributes                                                                      | categories
                        | city         | cool | date                        | funny | hours
                                                  | is_open | latitude | longitude | name                              | postal_code | review_count | stars_x | stars_y | state | text

                                                                                                                        | useful | user_id
------------------------+------------------------+--------------------------+---------------------------------------------------------------------------------+-------------------------
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

 bEeZg3fNNRCqg1v5DiSvXA | LYawlJRKGmOyLwpCZ0pCPQ | 819 Veterans Blvd, Ste 204 | {'ByAppointmentOnly': 'True', 'BusinessAcceptsCreditCards': 'True'} |
          Plumbing, Home Services |       Kenner |    1 | 2016-03-16 11:28:09.000000+0000 |     2 | {'Monday': '8:30-18:0', 'Tuesday': '8:30-18:0', 'Wednesday': '8:30-18:0', 'Thursday': '8:30-18:0', 'Friday
': '8:30-8:45', 'Saturday': '8:30-18:0', 'Sunday': '8:30-18:0'} |       1 | 30.00605 | -90.25821 | Eagle Plumbing Repairs & Drain Cleaning |       70062 |           38 |     1.5 |       5 |    LA |

                    At first i had an issue with the plumber and service, the manager called me back and was completely professional and courteous.  Manager sent out a new technician and helper. My problem
was solved in about 20 minutes to my complete satisfaction!  I highly recommend Eagle Plumbing for their professionalism, courtesy, and dedication to customer satisfaction! |      0 | wJDkJUixMETx10PsF4DKHg
 bEeZg3fNNRCqg1v5DiSvXA | M9y34Cf1csFwDUi23ahoCA | 819 Veterans Blvd, Ste 204 | {'ByAppointmentOnly': 'True', 'BusinessAcceptsCreditCards': 'True'} |
          Plumbing, Home Services |       Kenner |    0 | 2016-08-04 10:19:05.000000+0000 |     0 | {'Monday': '8:30-18:0', 'Tuesday': '8:30-18:0', 'Wednesday': '8:30-18:0', 'Thursday': '8:30-18:0', 'Friday
': '8:30-8:45', 'Saturday': '8:30-18:0', 'Sunday': '8:30-18:0'} |       1 | 30.00605 | -90.25821 | Eagle Plumbing Repairs & Drain Cleaning |       70062 |           38 |     1.5 |       5 |    LA |

                                                                                                                                                       Showed up on time. Pleasant and
 efficient. Took care of my shower plumbing issue and even adjusted the rain head in master bath!  Eagle Plumbing is definitely on my go-to list and should be on yours too! |      0 | k89tjTOMU_AhOToT7F4Opg
 bEeZg3fNNRCqg1v5DiSvXA | OCTn8JfS4lngrkHS0bqidQ | 819 Veterans Blvd, Ste 204 | {'ByAppointmentOnly': 'True', 'BusinessAcceptsCreditCards': 'True'} |
          Plumbing, Home Services |       Kenner |    0 | 2018-01-01 15:14:25.000000+0000 |     0 | {'Monday': '8:30-18:0', 'Tuesday': '8:30-18:0', 'Wednesday': '8:30-18:0', 'Thursday': '8:30-18:0', 'Friday
': '8:30-8:45', 'Saturday': '8:30-18:0', 'Sunday': '8:30-18:0'} |       1 | 30.00605 | -90.25821 | Eagle Plumbing Repairs & Drain Cleaning |       70062 |           38 |     1.5 |       5 |    LA |

                         Great Service came on a Sunday 12/31/17. Donald found the leak and fixed it! I was treading water and had four guests in my home. I was frantic. Price was reasonable and he
came the same day I called. Appreciate the promptness. He even had to go back to his shop to get a part to fix it. Five stars for Eagle Plumbing.\nLinda Warren\nNew Orleans |      0 | wfTxf_uyx0nK7T-SAaAg3A
 bEeZg3fNNRCqg1v5DiSvXA | x1rcrcNJfAPWNe2GdvkCxA | 819 Veterans Blvd, Ste 204 | {'ByAppointmentOnly': 'True', 'BusinessAcceptsCreditCards': 'True'} |
          Plumbing, Home Services |       Kenner |    0 | 2015-08-21 18:40:46.000000+0000 |     0 | {'Monday': '8:30-18:0', 'Tuesday': '8:30-18:0', 'Wednesday': '8:30-18:0', 'Thursday': '8:30-18:0', 'Friday
': '8:30-8:45', 'Saturday': '8:30-18:0', 'Sunday': '8:30-18:0'} |       1 | 30.00605 | -90.25821 | Eagle Plumbing Repairs & Drain Cleaning |       70062 |           38 |     1.5 |       5 |    LA |
```



```
Extension modules: cassandra.util, cassandra.cqltypes, cassandra.bytesio, cassandra.cython_utils, cassandra.deserializers, cassandra.parsing, cassandra.obj_parser, cassandra.row_parser, cassandra.protocol, ca
ssandra.connection, cassandra.cmurmur3, cassandra.query, cassandra.pool, cassandra.metadata, cassandra.io.libevwrapper, cassandra.cluster (total: 16)
Processed: 0 rows; Rate:       0 rows/s; Avg. rate:       0 rows/s
~7848210 rows imported from 1 files in 0 day, 0 hour, 3 minutes, and 23.876 seconds (0 skipped).
cqlsh:business_data> SELECT * FROM reviews LIMIT 10;

 business_id            | review_id              | address                | attributes                                                                     | categories                                  | city
            | cool | date                            | funny | hours                                                                                              | is_open | latitude | longit
ude | name                                  | postal_code | review_count | stars_x | stars_y | state | text

                                                                                                                                                                       | useful | user_id
------------------------+------------------------+------------------------+--------------------------------------------------------------------------------+---------------------------------------------+--------
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

 k8k-6ACN7-xVs6w91KJ3Xw | HqdVpRV1Kth89vvT_Q4R8Q |                   null | {'BusinessAcceptsBitcoin': 'False', 'BusinessAcceptsCreditCards': 'True'} | Event Planning & Services, Party & Event Planning |
Nashville |    0 | 2017-05-05 14:33:46.000000+0000 |     0 | {'Monday': '10:0-17:0', 'Tuesday': '9:0-17:0', 'Wednesday': '9:0-17:0', 'Thursday': '9:0-17:0', 'Friday': '9:0-16:0'} |       1 |    36.06 |    -86
.67 |                           NashVegas VIP |       37011 |            7 |       5 |       5 |    TN |

                                           Damon and his staff are easy to work with and reliable.  Planning a bachelor party from another state can be tough, especially since none of us had been
 to Nashville before, but Damon does a lot of research upfront to make sure your outing is exactly what you want it to be.  His hosts are accommodating and a lot of fun.  Nashville is a fantastic city and the
y know what to do and how to have a good time.  We came in Thursday afternoon and stayed through Sunday.  NashVegas planned the perfect amount of activities and hit a home run for us! |      1 | gBVodAmq8gMhy
-T6VVUbkA
 k8k-6ACN7-xVs6w91KJ3Xw | M6iZIR7uU3IVCUevwcWPvQ |                   null | {'BusinessAcceptsBitcoin': 'False', 'BusinessAcceptsCreditCards': 'True'} | Event Planning & Services, Party & Event Planning |
Nashville |    1 | 2015-08-07 14:35:39.000000+0000 |     0 | {'Monday': '10:0-17:0', 'Tuesday': '9:0-17:0', 'Wednesday': '9:0-17:0', 'Thursday': '9:0-17:0', 'Friday': '9:0-16:0'} |       1 |    36.06 |    -86
.67 |                           NashVegas VIP |       37011 |            7 |       5 |       5 |    TN |

                                          We used NashVegasVIP to plan a boat outing for our friend's bachelorette party.  We decided to try these guys because we were having a hard time organizing
 things from where we live in Atlanta.  We are all extremely happy we planned the boat day with them!!  It was exactly what we wanted, the price was reasonable, and everything was set up for us on the boat.
We had a blast on the lake!  The next time me and my friends get asked to host a bachelorette party, we will be using NashVegasVIP.  Without a doubt, I would recommend them to others. |      1 | NtxzjifiIZwuf
0H_KasQQQ
```





Datasets

- This is the business dataset, a sub-dataset of the Yelp Academic Dataset; for every type of business, this dataset shows restaurants, stores, service providers, and more.

- It includes important features such as star ratings given overall, review counts, business categories, and location details such as city, state, country.

- The data is both structured-some numerical fields comprise rating and review count-while other data is more categorically oriented, such as the type of business and its location. It is ideal for exploratory data analysis and machine learning.

- The dataset is in JSON format. Later, I convert it into CSV format and load it into PySpark for distributed processing, ensuring efficient handling and transformation of large volumes of business data.

- It contains information for several thousand enterprises and is thus a good basis for training machine learning models and doing predictive analytics on business success.
  - https://www.yelp.com/dataset

Conclusion & Future Work

- The project has effectively leveraged big data frameworks like PySpark, Dask, Cassandra, and Kafka in processing and analyzing large Yelp data. In addition, it has identified Random Forest and Gradient Boosting as the most appropriate predictions of business success.

- Future improvements include hyperparameter tuning, with the exploration of deep learning methods for further improvement in the accuracy of predictions and enabling more complex interactions between features.

- The extension of the system to real-time data processing using Apache Kafka and its integration with sentiment analysis using NLP will provide dynamic insights and a deep understanding of customer feedback.