**Report Summary**

Predicting the revenue of the restaurant is key metric indicator and plays a very vital role for making the future decisions. This Machine learning regression models try to build prediction system using various algorithms given the assignment.

This dataset has 137 restaurants in the training set, it has wide variety of restaurants to train the models. Revenue field is a revenue of the restaurant in the given year. This is what also called output value which we need predict for the given input unlabeled example given to predict for.

Reference to the dataset is taken from kaggle <https://www.kaggle.com/c/restaurant-revenue-prediction>

This dataset has test, train and sample submission datasets. However, for this assignment, I have considered only training dataset primarily as my focus was to build various machine learning models and did analysis of all the algorithms with their error logs, parameter tuning with various parameters depending on the each algorithms. In this document, I will explain each algorithm and their parameters.

This problem dataset has below fields which are considered useful for the training the models.

**City** : Restaurant city name where it is located.

**City Group**: What kind of city it is.

**Type**: Restaurant type. FC: Food Court, IL: Inline, DT: Drive Thru, MB: Mobile

**P1, P2 - P37**: There are three categories of these obfuscated data. Demographic data are gathered from third party providers with GIS systems. These include population in any given area, age and gender distribution, development scales. Real estate data mainly relate to the m2 of the location, front facade of the location, car park availability. Commercial data mainly include the existence of points of interest including schools, banks, other QSR operators.

**Revenue**: Revenue field is a revenue of the restaurant in the given year.

(Credit to Kaggle description: <https://www.kaggle.com/c/restaurant-revenue-prediction/data> )

**Tools I used**

**Python3**

This is one of the very popular language used across the machine learning community and developer as it has lot of support, tools and libraries. It was good learning and learned the importance of Python and its various libraries such as NumPy to do various numerical operation and collection operations. Used Matplot library to draw the graphs on the Jupyter. It has lot of options to draw lot of different graphs to see statistics.

**SparkML**

.This is the main library of the entire project. It is the main Machine Learning algorithms implementation library with lot of parameters to tune the performance of the models. This library supports both regression and also categorical algorithms. For this assignment I explored bit into regression algorithms as per the need basis.

**Jupyter**

This is also very useful and powerful tool to code, document and visualize the work we do to see the statistics, graphs, outputs and quickly experiment with various options. This is the easiest tool I learned from the exercise.

**Implementation**

Bike dataset example with documentation, code examples and explanation is really useful to understand the all various algorithms, what regression model, how to calculate error logs and their implementation. Learned a lot of concepts, examples and also most importantly learned how to implement a machine learning based logic and do performance tuning of the algorithm behaviour.

I built Bike dataset implementation by referring to provided assignment document. Functions provided in the document such as extract\_features, get\_mapping, extract\_label, extract\_features\_dt, squared\_log\_error, abs\_error and squared\_error are very helpful reusable functions so that easy to understand and implement my solution.

After finishing Bike dataset implementation, then I started with Restaurant implementation, implementation would be similar steps, however, main challenge was with finding the dataset, analyzing the all the fields in the dataset and read the dataset, cache the Resilient Distributed Datasets(RDD) and finally implement and analyse the all the given algorithms.

Dataset loading was one of the important task, and I spent some time here to learn how to choose and analyse.

|  |
| --- |
| path = "/regression-models/Restaurant\_train\_Remove\_Header.csv"  raw\_data = sc.textFile(path)  num\_data = raw\_data.count()  records = raw\_data.map(lambda x: x.split(","))  first = records.first()  print('First record: ', first)  print('Total number of records: ', num\_data) |

**Output**

|  |
| --- |
| Feature vector length for categorical features: 5 Feature vector length for numerical features: 37 Total feature vector length: 42 |

3 algorithms DecisionTree, Linear Regression and Gradient Boosted Tree were built as a part of the assignment.

As a part of the DecisionTree algorithm, implemented concept of Max Depth and Max Bins. One of the common thing to observe is the error rate between actual output and predictions made by the algorithms. This is the main key result to understand how the algorithm behaves for various options based on the algorithms and its parameters. Other main task is to find the various log implementations as per the implementation reference guide from the bike example such as Root Squared Error, Absolute Error and Mean Squared Error

As a part of the Gradient Boosted Tree, I implemented the parameters such as iterations, Max Depth and Max Bins.

As a part of the Linear regression, I implemented the parameters such as Intercept, Iterations, Step size, L1 Regularization and L2 Regularization.

All the outputs can be found the main documentation.

**Project Files**

Restaurant\_train.csv

This is the dataset I used for this assignment, it is taken from the Kaggle. I built the dataset loading, mappings and then build the various models for all the given algorithms.

RestaurantRegressionAnalysis.ipynb

Notebook is a complete example implementation of a Restaurant dataset. All the 3 algorithms implemented here in this notebook.

Restaurant-RevenueRegressionModels.doc

This is the main document to show the outputs of my code executions and graphs.

Restaurant-Revenue-Report

This one I used to show high level explanation of the assignment.

**How to run**

* Install Python3
* Install Spark

from <http://spark.apache.org/downloads.html>

Version: spark-2.3.0-bin-hadoop2.7

* Install Java 8
* Install Jupyter Notebook
* Configure following 3 System environment variables

export PYSPARK\_PYTHON = python3

export PYSPARK\_DRIVER\_PYTHON = jupyter

export PYSPARK\_DRIVER\_PYTHON\_OPTS = 'notebook’

* Run pyspark
* Opens Jupyter Home page in browser