Mini-Project 2: Yelp Business Rating Prediction using Tensorflow

Version 1.0

CSC215, Fall 2018

Oct 8th, 2018

Prepared by: Chandini Nagendra

Siddharth Chittora

Contents

[1. Problem Statement 2](#_Toc526176076)

[2. Methodology 3](#_Toc526176077)

[3. Experimental Results and Analysis 3](#_Toc526176078)

[4. Task Division 3](#_Toc526176079)

[4.1. Chandini Nagendra: 3](#_Toc526176080)

[4.2. Siddharth Chittora 3](#_Toc526176081)

[5. Project Reflection 3](#_Toc526176082)

[6. Additional Features 3](#_Toc526176083)

# Problem Statement

In this project, we aim to predict a business's stars rating using the reviews of that business and review count based on neural network implementation in Tensorflow. This project is twofold:

Task 1: Consider this problem as a regression problem. Compare the RMSE of the BEST **Tensorflow regression neural network model** you obtained with that of **regression model** you achieved in the last project.

Task 2: Consider this problem as a classification problem. Compare the accuracy of the BEST **Tensorflow classification neural network model** you obtained with that of **each classification model** you achieved in the last project.

# Methodology

Here we compare Linear and logistic models with Tensor flow models by using early stopping, Model checking and tuning the models with hyperparameters and see how they affect performance

## Regression Problem

## Linear Regression

**For 10000 Business:**

Root Mean Squared Error: 0.56

R2 score: 0.70

**For Full Data set:**

Root Mean Squared Error:

R2 score:

## Logistic Regression:

**For 10000 Business:**

Root Mean Squared Error: 1.38

R2 score: 0.54

**For Full Data set:**

Root Mean Squared Error:

R2 score:

## Tensor flow regression neural network models

#### Activation: ReLu

**Optimizer: adam**

##### Without stopping, checkpointing

**For 10000 Business:**

Root Mean Squared Error: 0.5732024908065796

R2 score: 0.68

**For Full Data set:**

Root Mean Squared Error:

R2 score:

##### With stopping, checkpointing

**For 10000 Business:**

Root Mean Squared Error: 0.5019386410713196

R2 score: 0.75

**For Full Data set:**

Root Mean Squared Error:

R2 score:

##### With additional features

**For 10000 Business:**

Root Mean Squared Error: 0.2876877188682556

R2 score: 0.92

**For Full Data set:**

Root Mean Squared Error:

R2 score

##### With stopping, checkpointing

**Optimizer: SGD**

**For 10000 Business:**

Root Mean Squared Error: 0.5619208812713623

R2 score: 0.69

**For Full Data set:**

Root Mean Squared Error:

R2 score:

**Optimizer: RMSprop**

**For 10000 Business:**

Root Mean Squared Error: 0.5563821196556091

R2 score: 0.70

**For Full Data set:**

Root Mean Squared Error:

R2 score:

**Optimizer: Adagrad**

**For 10000 Business:**

Root Mean Squared Error: 0.5838866233825684

R2 score: 0.67

**For Full Data set:**

Root Mean Squared Error:

R2 score:

**Optimizer: Adadelta**

**For 10000 Business:**

Root Mean Squared Error: 0.5558121800422668

R2 score: 0.70

**For Full Data set**:

Root Mean Squared Error:

R2 score:

**Optimizer: Adamax**

**For 10000 Business:**

Root Mean Squared Error: 0.5647493004798889

R2 score: 0.69

**For Full Data set:**

Root Mean Squared Error:

R2 score:

**Optimizer: Nadam**

For 10000 Business:

Root Mean Squared Error: 0.5831780433654785

R2 score: 0.67

**For Full Data set:**

Root Mean Squared Error:

R2 score:

From these results adam optimizer performed the best, so we use that to continue our trials

With 2 hidden layers

**For 10000 Business:**

Root Mean Squared Error: 0.6047360897064209

R2 score: 0.65

**For Full Data set:**

Root Mean Squared Error:

R2 score:

With 3 hidden layers

**For 10000 Business:**

Root Mean Squared Error: 0.59750896692276

R2 score: 0.65

**For Full Data set:**

Root Mean Squared Error:

R2 score:

With 4 hidden layers

**For 10000 Business:**

Root Mean Squared Error: 0.5962139368057251

R2 score: 0.66

**For Full Data set:**

Root Mean Squared Error:

R2 score:

With 5 hidden layers

**For 10000 Business:**

Root Mean Squared Error: 0.5959509611129761

R2 score: 0.66

**For Full Data set:**

Root Mean Squared Error:

R2 score:

## Classification Problem

### KNN

### SVM

### Naïve Bayes

Tensor flow

# Experimental Results and Analysis

# Task Division

## Chandini Nagendra:

## Siddharth Chittora

Discussed together on how to improve the model and came up with the following solution

# Project Reflection

label encoding does not work for tensor flow, when using classification. do one hot coding

using the same best weight model had interfered with the model performance - elaborate

used zscore for tanh as the range is from -1 to +1

For smaller data set the activation function sigmoid gave the least RMSE or better result

# Additional Features

for reviews, review count

L1 Regularization

Dropout