MSA Phase 2 Executive Summaries.

Part 1: Analysis and Preprocessing

The data provided is for weekly sales in W stores and provided to use across three files. Overall we were provided weekly sales data as well as ID information for stores and departments within. Additional feature information about the store and environment nearby was also provided. All variables provided are:

- Store, Dept, Date, Weekly_Sales, IsHoliday, Temperature, Fuel_Price, MarkDown1, MarkDown2, MarkDown3, MarkDown4, MarkDown5, CPI, Unemployment, Type and Size

Our exploration of the sales data set and visualisation of the different features show that only a few features influence the weekly sales variable, namely Size, Deptmartment and Type. The features Temperature, CPI, Unemployment and Fuel_Price did not have a visible influence on Weekly_Sales and could be dropped. The IsHoliday feature, although only 4 weeks of the year, could be seen to boost sales, especially around the end of each year. The variables MarkDown 1 to 5 showed a weak correalation with weekly sales, however over 60% of data was missing and we may not be seeing the full picture. To handle these missing values three new dataframe were prepared. One in which null values were dropped completely, however in doing so valuable information of other variables could have been lost. A second one in which the null values were filled with zeros, and the third replaced the null values with an average over the weeks to capture what it might have been. Each dataframe was saved to a csv file to be imported into part 2.

Part 2: Training and Evaluation

Part 2 used the result from part 1 to train a regression model and make predictions about future weekly sales. For this part we used the saved dataframe in which null values for MarkDown's were filled with zeroes. We tested 6 different regression models, these were Random Forest Regressor, Linear Regression, K Neighbours Regressor (KNN), Decision Tree Regressor, Gradient Boosting Regressor, Ridge. Each model was trained on a 70% data split and test using the remaining 30%. Our results show that Random Forest Regressor had the overall best performance with the highest r2-score and lowest root mean square errors. Decision Tree Regressor was a close second. Random Forest Regressor parameters were then tuned and the model's performance was plotted. We find the model performs very well. Future predictions were made for one year following the end of the provided sales data and results plotted against past years. Our model predicts future sales with a similar series as past data. It also accounts for sales growth around the holiday periods.

Part 3: Deep Learning

For this image classification task a convolutional neural network was implemented. The image set was split into a test-validation-test split at a 70-20-10 ratio. Each image was normalised to a [0,1] range and had a random horizontal flip transformation applied. For training we implemented a training loop and a validation loop saving a number of metrics to evaluate our model at each epoch, including accuracy, precision, recall, f1-score and loss. An early stopping was also implemented to stop our training if the validation loss no longer improved, the model with best validation loss was saved.

Following our training metrics for the training and validation loops over each epoch were plotted and showed that training losses decreased and accuracy increased with epoch numbers. However, although the validation curve had the same general shape as the loss there is a lot of fluctuation, and the model is not generalising well, an indication of overfitting. With the saved model the test split was used to further evaluate the trained model the evaluation metrics show our model predicted ~55% of the test split images overall. Plotting a confusion matrix gave insight into which images the model had trouble with labelling 5 and 0 as well as 9 and 8. Finally, the model was to predict the images given to us in the test folder, those without labels and the result was written to a csv file for submission.

Overall the model demonstrated an average performance. To try address the possibility of overfitting we did apply a number of dropout layers, but additional optimisation/regularisation techniques will help to improve performance of the model further.