**Delivery 2: EDA and Similar Researches**

- David Gill *(Data 606)*

**METHODOLOGY:**

1. Csv files were saved on my local machine
2. Python was used to perform data cleansing and exploratory data analysis

**EDA RESULTS:**

1. Stores file contains 45 Rows 🡪 which means there are 45 stores
2. It contains 3 Columns/Attributes - **Store, Type, Size**
3. There are 3 distinct Store Types - **A, B, C,** with **A** beingthe largest and **C** the smallest
4. There is no overlapped area in the sizes of the stores
5. There is no missing data in this file
6. Features file contains 8190 rows and 12 attributes, showing the features/factors affecting sales
7. Train data contains 421570 records and test data contains 115064 records
8. Sales on holidays was found to be a little bit more compared to non-holidays
9. The Department with highest sales lies b/w 60 – 80

**SIMILAR RESEARCHES/ PROJECTS:**

**1)Electronic Devices Sales Prediction Using Social Media Sentiment Analysis –**

<http://cs229.stanford.edu/proj2012/ZarghamNassirpourNasiri-ElectronicDevicesSalesPredictionUsingSocialMediaSentimentAnalysis.pdf>

- In this project they predicted the sales of electronic devices based on the sentiment of the comments made about the products, before their release, on Twitter. Data used was pertaining to social media content and product sales. The sentiments of the comments were predicted using a machine learning framework based on recursive autoencoders (RAE) for sentence-level prediction of sentiment label distributions *(semi-supervised)*, further using 70/30 cross-validation on this data, settling on classification accuracy of 83%. They, then used linear regression with four features to predict sales of the products based on the sentiment analysis.

**2) Predicting sales in a food store department using machine learning**

<http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1108597&dswid=5794>

-Their study aimed to compare three machine learning methods for sales prediction in the food industry - Multilayer Perceptron (MLP), Support Vector Machine (SVM) and Radial Basis Function Network (RBFN). The performance of the models was determined using the performance measures: Mean Average Percentage Error (MAPE) and Root Mean Squared Error (RMSE). Based on the results, the SVM performed with lower error measures than the other two methods and was concluded to be the best. The data consisted of sales data provided by a Swedish food company; pertaining to one department in one store from year 2012 to year 2016. Each validation set then consists of 180 daily sales in a department.

**3)Drugs store sales forecast using Machine Learning** <https://s3.amazonaws.com/academia.edu.documents/59368319/191_report20190523-80443-dzybc.pdf?response-content-disposition=inline%3B%20filename%3DDrugs_store_sales_forecast_using_Machine.pdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAIWOWYYGZ2Y53UL3A%2F20200302%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Date=20200302T021231Z&X-Amz-Expires=3600&X-Amz-SignedHeaders=host&X-Amz-Signature=07f88ef3f10d237291a4ae68c4935291b9f6de8be7cbb1dae1a90452aad42c24>

- For this project, training data of 1115 of Rossmann stores’ daily sales dated back to 2013, with 1,017,209 entries in total, including features of promotion and competitors’ information was used. Since they had no access to the real sales amount for testing during Kaggle competition, so they used 70% of the contest given training data as the training set for their model, the rest 30% as test set for cross validation. They established an auto regression (AR) model and tested it using order numbers and calculated the test errors. Random forest (RF) and Support regression vector (SVR) were used, to help identify the most apt feature/factor influencing sales. They made good predictions based on the adoption of the above-mentioned models.

**OBSERVATIONS:**

1. Data used is pertaining to sales, including features/factors affecting it. I am working on a similar dataset. The goal of their project is like mine – predicting sales, only the domain is a little different
2. They have used a variety of models for testing dependency of features and the accuracy of the model. I am planning to use Random Forest and /or extra trees for sales prediction
3. In my EDA I have analyzed the type, sizes of stores, frequency of sales for each department and holidays affecting the sales
4. Their projects have used a method to verify accuracy of their models
5. I will proceed to merge and train the data to understand the features affecting sales. I must predict sales for 45 stores based on the departments.