**Overdose AI**

**Data Scientist Intern Practical Exercise**

**REPORT**

**G Ravindernath yadav**

[**Ggrnathyadav2960@gmail.com**](mailto:Ggrnathyadav2960@gmail.com)

**For the ease of understanding the report is divided into the following sections**

* Introduction to the data
* Questions for the data and their subsequent answers
* Methodology of the analysis
* Key insights
* model preparation and results

**INTRODUCTION TO THE DATA**

The data has 18000+ rows divided into 14 columns. The data need to be cleaned and the data types will have to be altered.

**QUESTIONS FOR THE DATA?**

* Should the missing values be dropped?
* Initial correlation matrix revels no insights and hence we must do feature engineering to make sense of correlation matrix
* The data type of certain columns needs to be change. For example, columns containing date should be of datetime data type
* Is there a relation between the status and price, item type and price, physical dimension and price
* Removal of outliers

**METHODOLOGY OF ANALYSIS**

The data was cleaned and rid of outliers. There after we did feature engineering to column width and thickness and combined them into a new column named “ l\_w”(sorry for bad naming). Furthermore we needed to see how price varies with respect to date, day , month and year which was achieved through feature engineering of the item\_date , delivery date columns into their respective formats. After this we weighed in our main features based in the insights gathered.

**INSIGHTS**

There is a strong correlation between the “l\_w” column and selling price and the year of delivery and month of item(i\_month). The columns application and prodect\_ref too have a higher correlation compared to the other columns and hence these columns should be chosen as features for over model to fit .

**MODEL PREP AND RESULTS**

The data was divided into training data and testing data (90:10), and the features where scaled using StandardScaler(). Models chosen for preliminary testing are as follows:

* LinearRegression,
* Lasso,
* ElasticNet,
* DecisionTreeRegressor,
* KNeighborsRegressor,
* GradientBoostingRegressor

The models were fit and the resulting r2scores were analyzed. Models DecisionTreeRegressor, KNeighborsRegressor, GradientBoostingRegressor all had a better r2score than the simple linear regression which makes sense giving the polynomial nature of the regression.