**Guess the product report**

1. **Exploratory Data Analysis of variables :**

**Document name – guess\_the\_product\_eda.ipynb**

1. Average number of words used per product description is good. ( more than 60 )
2. There is a high bias of total number of records per product.

Around 74 % records of top 5 products. (1758, 1274, 1522, 1250, 1376).

Wordclouds of each of this companies visibly gives us an idea about the sectors they must have active in: Transport, Manufacturing, Real estate, Resource allocation and management, etc.

1. Average billing amount is around 40 ( wrt. currency )
2. Text description is playing very important role in diversifying product categories.
3. Sensible trailing of words in n-grams has been found.
4. **ML Model :**

**Document name – guess\_the\_product\_ML.ipynb**

1. **Data Transformation –**

Removed Invoice id as an irrelevant field.

Label-Encoding for Product category and Vendor code.

NLP cleanups on Description text data.

1. **NLP –**

Using Count-Vectorizer to get the sparse description vector.

TF-IDF to highlight impactful words and normalize weights of common words across the document.

Corpus size after cleanup – 89,013

Unique words in corpus – 2,557

n-grams: Store Management, Corporate service, Car field,

Travel Entertainment

Can be effectively used for semantic embedding’s to model.

1. **Naïve Bayes Classifier –**

For the given set of data considering complexity levels, strong independence assumptions between the features of data. Naïve Bayes classifier works best for fitting on such data.

1. **Model Evaluation :**

Using Multiclass metric to evaluate the current task.

Represents accuracy based on prediction, label, weight (optional) and probability (only for log-Loss).

Accuracy: ~ 90 %

1. **Submission file :**

Inverse Transformed the Targets.

Exported DataFrame in required format.

1. **Improvements :**
2. **Data can be cleaned further by maintaining custom noise list.**
3. **Date parameter have been observed and can be used as a feature to check the relationship of demand with respect to seasons.**
4. **Model accuracy can be further improved by tuning the hyper-parameters.**

**Or using advanced algorithms such as,**

**Extreme gradient boosting.**

1. **Ensemble techniques can be performed to get improved results.**