Product Length Prediction

Lab Project submission submitted for

Machine Learning (UML501)

submitted by

Rushil Agarwal(102103757)

Nischay Morya (102103763)

3 COE 27

submitted to

Dr. Ashutosh Aggarwal

Computer Science and Engineering Department

Thapar Institute of Engineering and Technology, Patiala

Introduction:

In today's day and age ecommerce websites have become an integral part of our lives. Be it a student or a businessman, everyone in some capacity interacts with such websites on a daily basis. These websites are filled with endless products, each seeming more attractive than the other. However, sometimes customers are misled due to misinformation or no information, and the product dimension is one of them.

Our project aims to rectify this problem using machine learning techniques on the dataset provided during Amazon ML Challenge 2023 which provides us with a vast amount of data to work with.

Using this data we aim to predict the length of a product which not only makes the customer make a better decision but helps us in further deciding the packaging and shipping requirements of a product.

Significance of the project:

The primary objective of this project is to build a robust machine-learning model to predict the product length from the Title, Description and Bullet Points of a product which are provided by the seller at the time of listing the product.

This project can be deployed on any of the e-commerce websites and provide the sellers with the option to recheck their product dimensions in cases of discrepancy and auto-suggest an estimated dimension in cases of missing information so that there is no confusion when buying the product for the customer.

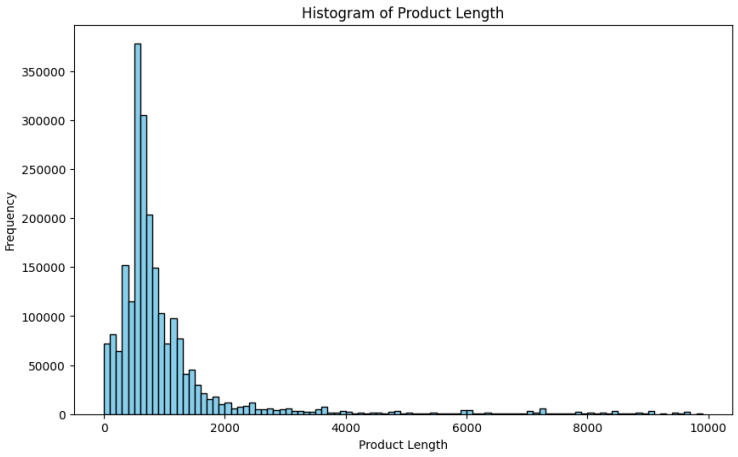
We can also use the dimensions provided by the model to pick the size of the packaging required for the object to be shipped and also how can the product be shipped for e.g. it will help us identify if the product can be delivered on a 2-wheeler or a bigger transport might be required for it even before it leaves the seller's warehouse! Although very niche, this will help us in further optimization of delivery vehicles and will probably help in the long run in saving time, money and fuel for the company.

This problem is relatively new and under-researched. The data is raw and widely varying, and this is our attempt at tackling it.

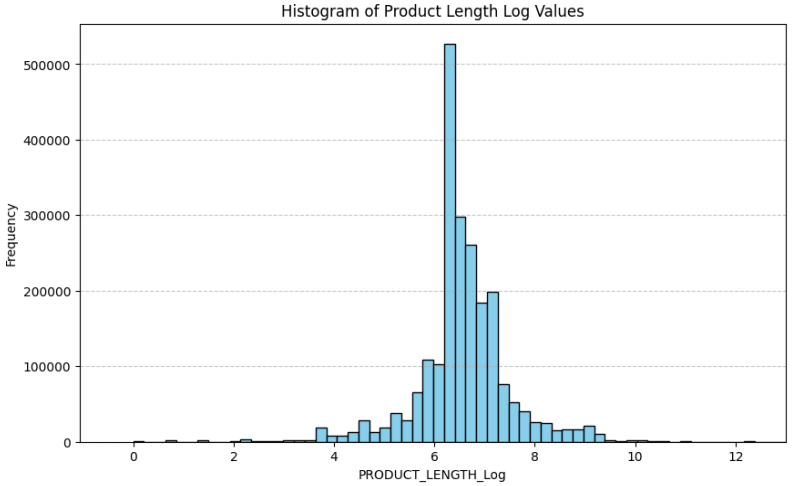
Methods Used:

Pre-Processing:-

1. First, we transformed the text using lowercase and removing alphanumeric characters and removing all the stop words (a, an, the) as they don't help with embeddings.
2. Then we concatenated all the text features (Product Title, Description, Bullet Points) into one to be embedded using Word2Vec.
3. The data was highly positively skewed, so we took the logarithm of the target values (Product Length) to fix skewness and bring it to a Gaussian shape.
4. We clipped the logarithm of the target values at 12 to remove outliers.



Data before preprocessing (Positive Skew)



Data after preprocessing (Gaussian)

Word2vec:-

1. Using word to vec we represent every word as a vector dimension 100 as it is a good middle ground and helps us find general relations between words without costing us much computing power.
2. We use the ‘workers’ parameter to make use of multiple cores in our CPU for faster computation
3. After this, we save the model for further usage

ANN:-

1. Used a shallow artificial neural network with 100 neurons in the input layer
2. Mean square error is used as loss metric
3. 50 epochs used with a batch size of 32.
4. It outputs a singular value corresponding to the product length in logarithmic scale.

Linear Regression:-

1. Using standard linear regression we form a relationship between the embeddings and the product length that we have to predict. We convert the embeddings into a numpy array and use them as features and use product length as the target to predict.

Testing:-

1. We use our model to run tests on 2 different groups of test data that we have designed.
2. For ANN:-
    a. For our first test case we take the bottom 100k entries of the data we used to train our model. We re-generate embeddings for these just in case and then predict values after that we calculate the error of every entry and then take an average which comes out to be 29.91625% this gives us an accuracy of 70.08375%.
    b. For our 2nd test case, we use the immediate 100k entries which are not present in our data. We generate embeddings similar to the first test case and find out that error and accuracy are 30.20707 and 69.79292 respectively.
3. Linear Regression:-
    a. Similarly to ANN we Embed our data (100k entries) and then predict values for both the dataset and find error and accuracy of our predictions made.

Considering the submissions for the competition and the dataset which we were able to use we can say that these predictions are satisfactory.

Dataset Description:

Our dataset consists of 6 attributes and 22,49,698 instances.

The attributes are as follows:

1) PRODUCT_ID: A numeric ID that uniquely identifies a product listing on the website.
2) TITLE: The name of the product that is displayed on the website. Contains brief information and basic specifications of the product.
3) BULLET_POINTS: Contains major points of interest that identify the product.
4) DESCRIPTION: Complete and detailed description of the product containing all the information that the seller can provide.
5) PRODUCT_TYPE_ID: A numeric value that identifies the category that the product belongs to.
6) PRODUCT_LENGTH: This is the target variable. It contains the length of the product in millimetres (mm).

After analyzing the dataset we decided to use the Title, Bullet_Points and Description attributes in conjunction to deduce the target variable i.e. Product_Length.

Data Columns:-

```
train_data_columns = train_data.columns
print(train_data_columns)
```

```
Index(['PRODUCT_ID', 'TITLE', 'BULLET_POINTS', 'DESCRIPTION',
       'PRODUCT_TYPE_ID', 'PRODUCT_LENGTH'],
      dtype='object')
```

Data Sample:-

	PRODUCT_ID	TITLE	BULLET_POINTS	DESCRIPTION	PRODUCT_TYPE_ID	PRODUCT_LENGTH	P
500000	1273981	BTween Big Girls Pleated Dress with Pearl Neck...	[Pleated chiffon skirt,Back zipper,Detachable ...	NaN	2738	1090.0	
500001	355053	Devil in Winter (Wallflower)	NaN	NaN	3385	675.0	
500002	576577	Ashrae Greenguide: An Ashrae Publication Addre...	NaN	NaN	140	825.0	
500003	949762	Nostalgic Warehouse Rope Keyhole Cover, Oil-Ru...	[Made of solid forged brass,Measuring 2&#x2033...	A classic rope border makes this oval keyhole ...	10200	200.0	
500004	1720924	Christmas Red Plaid Faux Fur Lighted House 6 i...	[Brighten up your home with this festive home ...	RAZ Imports is a wholesale importer of Seasona...	8490	400.0	

Experimental results:

ANN Results:-

1st Dataset

```
Error of values : 29.91625473248112
Accuracy of values: 70.08374526751888
```

2nd Dataset

```
Error of values : 30.2070744893559
Accuracy of values: 69.7929255106441
```

Linear Regression Results:-

1st Dataset

```
Error of values : 32.13777526118367
Accuracy of values: 67.86222473881634
```

2nd Dataset

```
Error of values : 39.33149801927422
Accuracy of values: 60.66850198072578
```

Final Results:-

		Dataset	
		Dataset 1	Dataset 2
Error	ANN	29.91625473	30.20707449
	Linear Regression	32.13777526	39.33149802
Accuracy	ANN	70.08374527	69.79292551
	Linear Regression	67.86222474	60.66850198

Individual roles:

1) Rushil Agarwal
    - Handled the data exploration and preprocessing
    - Performed tokenization and embedding

2) Nischay Morya
    - Applied ANN and Regression to the embeddings
    - Created tests and implemented them for the models used to find errors and accuracy.