Ahmedabad
University

CSE523 Machine Learning
Prof. Mehul Raval

Authors
[1]Rishabh Misra
Mengting Wan
Julian McAuley

Abstract— **Product size recommendation and fit prediction are critical in order to improve customers' shopping experiences and to reduce product return rates. This machine learning project aims to predict the fit of a cloth as fit, small or large, based on various parameters such as age, weight, height, body type, size, reviews, and ratings. The objective of this project is to develop an accurate and reliable prediction model that can help customers make informed decisions while purchasing clothes online. The model will be trained on this dataset using multiple machine learning algorithms such as linear regression, logistic regression, decision trees, and random forests. The ultimate goal of this project is to provide personalized recommendations to customers based on their body type, size, and preferences, thereby improving their online shopping experience.**

Keywords— **Logistic Regression, Size Prediction Model, Correlation, Labels, Classification., Exploratory Data Analysis**

I. INTRODUCTION

Cloth size prediction is an interesting and useful machine learning project that involves building a model that can predict the size of clothing that would fit a person based on their body measurements. The goal of this project is to help people find clothes that fit them better, and reduce the frustration of buying clothes that don't fit. The project involves collecting data on body measurements and clothing sizes, and using this data to train a machine learning model. Then we will perform Exploratory Data Analysis. The model will then be able to take in new body measurement data and predict the appropriate clothing size for that person. To build a cloth size prediction model, we will first need to collect a dataset of body measurements and corresponding clothing sizes. Once we have our dataset, we will need to preprocess the data by cleaning it, removing any missing values, and scaling the features. We will then split the data into training and testing sets and use the training set to train our machine learning model. There are several machine learning algorithms that we can use for this project, including logistic linear regression, random forests, and . We will need to evaluate the performance of each algorithm and select the best one for our project. Finally, we will use the trained machine learning model to make predictions on new data and evaluate its performance on the testing set.

II. LITERATURE SURVEY

Only a few researches have been proposed so far, and the product size recommendation problem is still relatively new. One method recovers the actual sizes of the products and consumers and uses them as features in a common classification method for fit prediction. Many researches used different machine learning models like KNN algorithm, random forest regression for predicting cloth size prediction model. Existing researches suggest logistic regression models with ordinal categories to improve model fit. It takes a different approach where focus is on capturing fit semantics and use correlation matrix methodologies with prototyping to address label imbalance concerns. Then, logistic regression has been applied on the model to find out the accuracy of predicting cloth size prediction. Another recent method learns the latent properties of customers and products using skip-gram models. This method, however, presupposes the existence of platform-specific properties, whereas the previous model operates on more constrained problem.

III. RESULTS

There were many null values in the data set so first we remove all the null values from the data set. We have converted height from feet to inches, and converted data types of variables height and weight into integers. Then we perform Encoding on a variable and convert it to integer data type. After that we perform EDA and find out correlation among variables like fit user_weight, user_height, product_size and user_age and make a correlation matrix. Through this we have generated a heat map. And through the box and whiskers plot we have found outliers in the given variables. The final result of Logistic Regression came out to be an accuracy score of 68%.

IV. IMPLEMENTATION

The implementation has been divided into the following stages:

1. **Finalizing the data set:** Out of two datasets, we checked the fields and null values of both. It was discovered that one of the datasets had more numbers of null values than others by a significant margin. Thus, we decided to continue with the second dataset as it gave us a better understanding of our problem statement.

2. **Checking null values:** Next stage was to check null values. Since a dataset of 1,90,000 values had a maximum of 10,000 null values, we decided to drop them. We also performed data imputation using standard deviation to scale the target feature.

3. **Data cleaning:** All the fields of the dataset were objects initially. We found some of them have to be converted to either float or int for Exploratory Data Analysis. The next step was to give a suitable field name to each column. We also created functions to convert a height object into a height in inches.

4. **Encoding**: The following stage was to apply nominal and ordinal encoding on categorical data. We concluded that five columns of prime importance needed to be encoded- fit, bust size, body type, rented for, and product category. We used ordinal encoding on fit and bust size. OneHotEncoding was used on body type, rented for, and product category.

5. **EDA:** Next stage was to find patterns and make a raw hypothesis on the dataset. We performed exploratory data analysis using a heat map, correlation matrices, histograms, and box and whiskers plots to do the same. It helped us study our data from multiple dimensions.

6. **Applying regression model:** Finally, after understanding the data and learning new patterns, we applied a regression model to the data. Through Logistic Regression, we tried to predict the accuracy of the model.

V. CONCLUSION

1. After performing data cleaning, exploratory data analysis and encoding, we applied a machine learning algorithm to predict our target feature.
2. We applied a logistic regression model to predict product size. Our model achieved an accuracy of 68%, which indicates that it performs better than chance. However, there is still room for improvement in the model's performance.
3. Overall, our study highlights the importance of careful data preparation and exploratory data analysis in machine learning projects. Further research can be done to improve the accuracy of the model, such as trying different algorithms, feature engineering, or optimizing hyperparameters.

VI. REFERENCE

[1] **Clothing Fit Dataset for Size Recommendation. (2018, August 21). Kaggle. https://www.kaggle.com/datasets/rmisra/clothing-fit-dataset-for-size-recommendation.**

[2] **Misra, Rishabh, Mengting Wan, and Julian McAuley. "Decomposing fit semantics for product size recommendation in metric spaces." In Proceedings of the 12th ACM Conference on Recommender Systems, pp. 422-426. 2018.**

[3] **Misra, Rishabh and Jigyasa Grover. "Sculpting Data for ML: The first act of Machine Learning." ISBN 9798585463570 (2021).**

[4] **Vivek Sembium, Rajeev Rastogi, Atul Saroop, and Srujana Merugu. 2017. Recommending Product Sizes to Customers. In RecSys.**

[5] ***CampusX. (2021, April 13). One Hot Encoding | Handling Categorical Data | Day 27 | 100 Days of Machine Learning [Video]. YouTube. https://www.youtube.com/watch?v=U5oCv3JKWKA***