## **ML Mini Project Report**

### **Introduction**

Machine learning plays a pivotal role in analyzing user interactions with online platforms. This project focuses on analyzing impressions—a representation of user interactions with products—to gain insights into user behavior and product performance. The dataset used consists of 965,003 unique rows, with each row capturing critical features related to users, products, and their interactions. All models for this project were built from scratch, and placeholders for code snippets are provided throughout this report.

The dataset used for this project is the FairJob Dataset, which contains anonymized job advertisement impressions and user interactions, aimed at exploring fairness and personalization in recommendation systems.

### **Features and Labels**

**Features**:

* **user\_id**: A unique anonymized identifier for each user.
* **product\_id**: A unique identifier for each product (e.g., job offers).
* **impression\_id**: A unique identifier for each impression (i.e., online session involving multiple products).
* **cat0 to cat5**: Anonymized categorical user features.
* **cat6 to cat12**: Anonymized categorical product features.
* **num13 to num47**: Anonymized numerical user features.

**Labels**:

* **protected\_attribute**: A binary feature indicating user gender proxy:
  + 0: Female
  + 1: Male
* **senior**: A binary feature indicating job seniority:
  + 0: Assistant role
  + 1: Managerial role
  + Derived from product titles, with keywords like ‘President’ or ‘CEO’ indicating a managerial role.
* **rank**: Positional rank of the product on the display for a given impression. Lower rank corresponds to a higher position on the display, often influencing click behavior.
* **displayrandom**: A binary feature indicating whether the product’s position on the banner was randomized (1) or not (0). Click-rank metrics are computed for randomized displays to minimize positional bias.
* **click**: A binary feature indicating whether the user clicked on the product during the impression (1 for clicked, 0 for not clicked).

### **Data Overview**

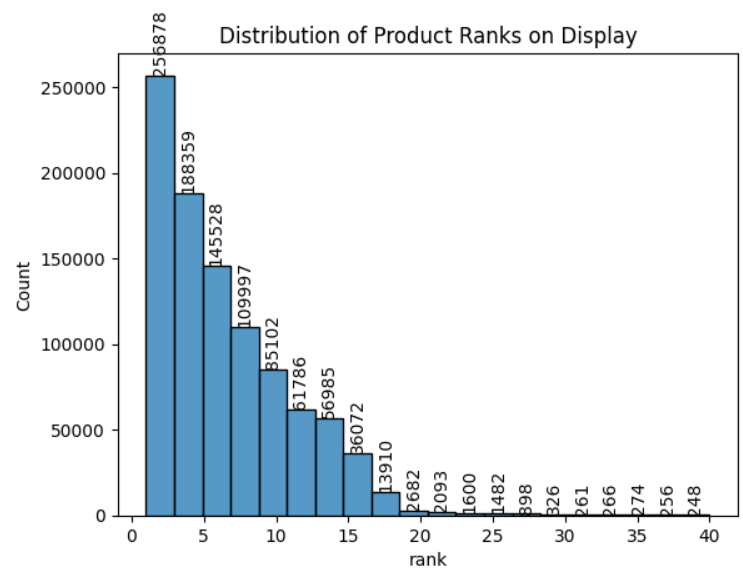
* **Total Rows in Train Set**: 965,003
* **Uniqueness**: All rows are unique.
* **Unique Identifiers**:
  + **User ID**: 29,823 unique values
  + **Product ID**: 55,845 unique values
  + **Impression ID**: 218,278 unique values
* **Key Relationships**:
  + Only 6,447 rows have repeated combinations of *'product\_id'* and *'impression\_id'* features.
  + Total rows with repeated keys: 12,894.

**Key Insight**: The combination of *'product\_id'* and *'impression\_id'* can be considered a primary key, representing a unique impression for a given product.

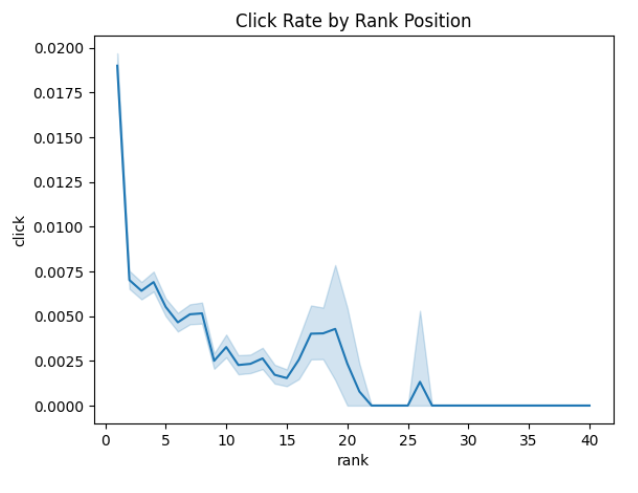
**Definition**: An impression refers to a user’s interaction with a product during an online session.

### **Click Behavior Analysis**

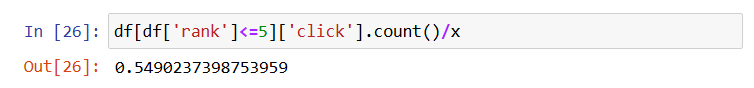
**Based on Rank**:



* Distribution of product ranks.

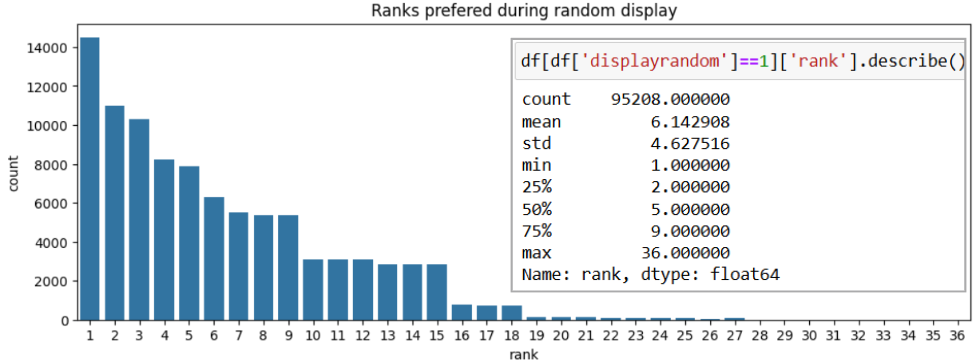


* Higher rank positions correlate with a higher click rate.
* Click rates rise between ranks 15 and 20, likely due to pagination on job portals.
* Ranks below 26 showed no clicks, indicating a sharp decline in visibility.

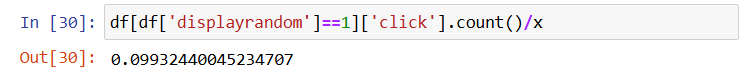


* Over 50% of the products are displayed at ranks ≤5, emphasizing strategic product placement.

**Based on Random Display**:



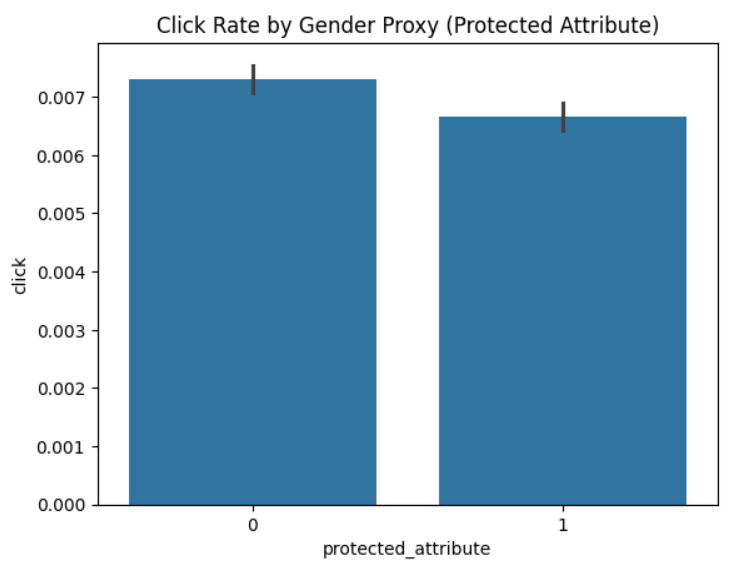
* When ads are displayed randomly:
  + Over 50% appear at ranks ≤5.
  + 75% appear at ranks ≤9.



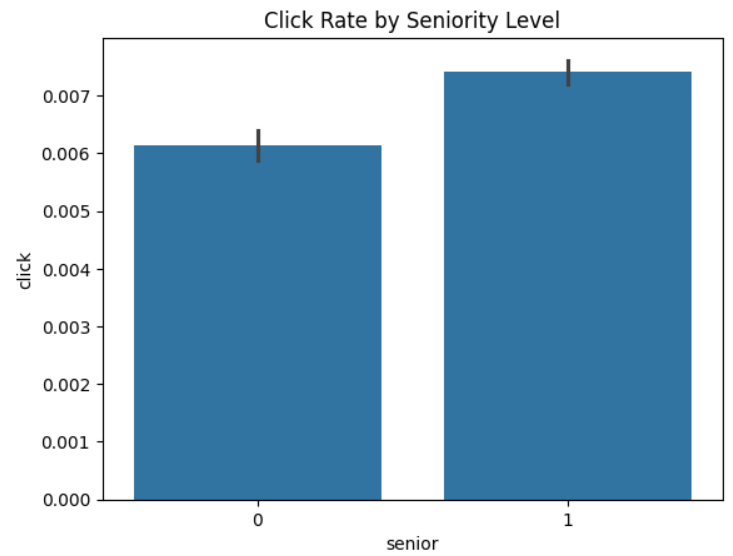
* Approximately 10% of impressions involve random display.

**Feature vs. Click Relationships**:

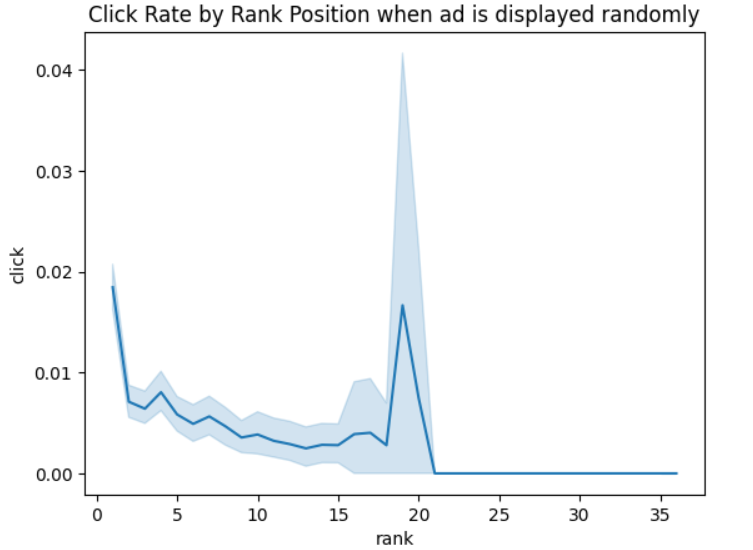
* Only 0.703% of products are clicked, highlighting the importance of a stratified split on the ‘click’ feature.



* Females click more frequently than males.



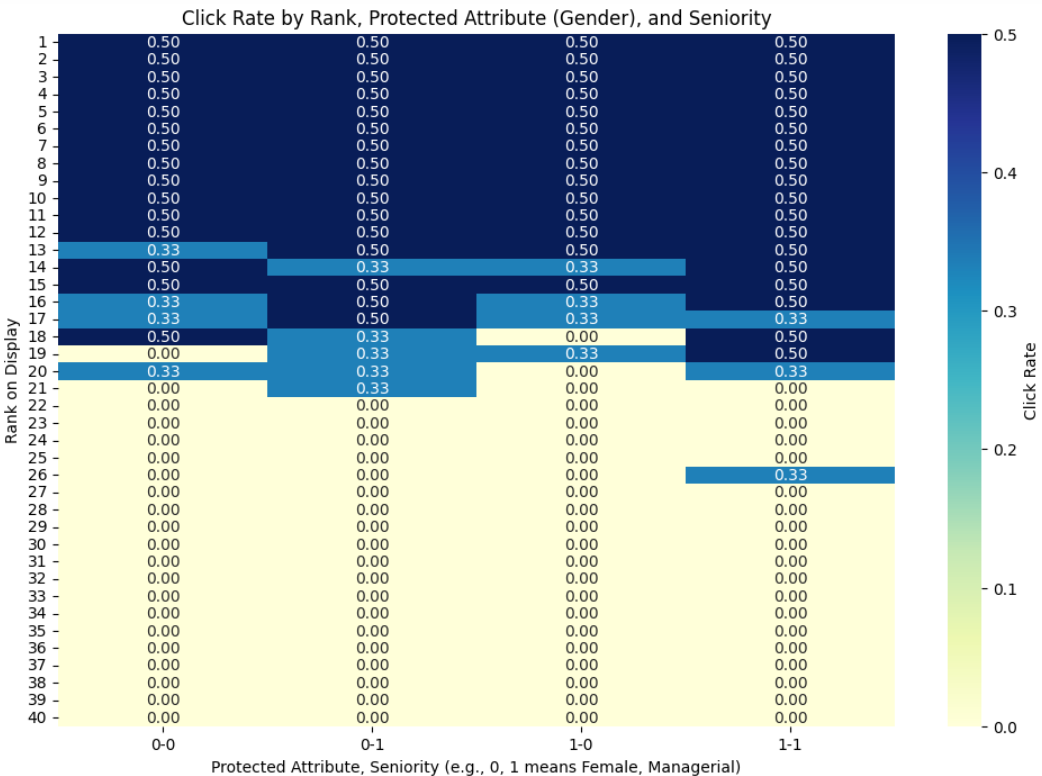
* Managerial roles receive more clicks compared to assistant roles.



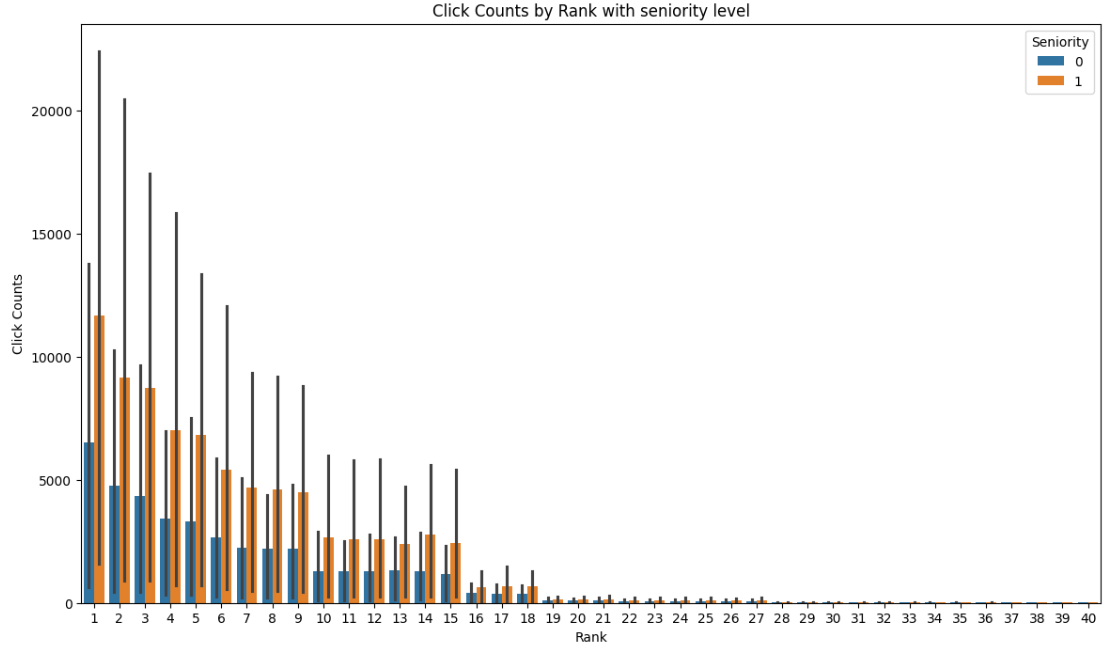
* When displays are random, ranks 18 and 19 have click rates comparable to rank 1.

### **Gender-Specific Insights**

**Gender and Role Preferences**:



* **Females**:
  + Prefer assistant roles more than males.
  + Are highly likely to click on lower-ranked ads.
* **Males**:
  + Show higher preference for managerial roles.
  + Are equally likely as females to click on lower-ranked ads.



**Overall Trend**: Regardless of rank, managerial roles are more likely to be clicked than assistant roles.

### **Feature Analysis**

**Cosine Similarity**:

* **User Features (cat0-cat5) and Product Features (cat6-cat12)**:
  + High similarity (near 1) across splits indicates stable Principal Component Analysis (PCA) projections.
* **Numerical Features (num13-num47)**:
  + Lower similarity (near 0.56) suggests variability or noise, requiring preprocessing refinement.

**Challenges**:

* Missing data in ‘above\_10’ and ‘below\_10’ splits necessitates debugging.
* Ensuring uniform preprocessing and PCA application across splits is crucial for consistent results.

### **Performing Classification**

Since we wanted to know the percentages of confidence and perform conditional probabilities, we decided to use Logistic Regression and Naive Bayes respectively.

We have defined them in classes and imported them into the analysis codes. For both techniques, we have used accuracy, recall, and F1 scores as the metrics.

#### **Logistic Regression**

**Model Design**:

* A softmax function was used as the activation function to transform raw scores into probability distributions across multiple classes.
* The model parameters (weights and biases) were initialized to zero and iteratively updated using gradient descent.
* The loss function minimized during training was categorical cross-entropy.

**Training Process**:

* Input features (‘X’) were used to compute the linear combinations of features and weights, followed by the softmax transformation for probability predictions.
* One-hot encoding was applied to the target labels (‘y’) to compute gradients for each class.
* Gradients of the loss function with respect to weights and biases were calculated and used to update parameters iteratively.

**Evaluation Metrics**:

* **Accuracy**: Proportion of correctly classified samples.
* **Recall**: Ability of the classifier to identify relevant instances correctly.
* **F1 Score**: Harmonic mean of precision and recall, balancing performance across imbalanced datasets.

#### **Naive Bayes**

**Model Design**:

* The prior probabilities of each class (‘class\_priors’) were calculated based on class frequencies.
* Conditional probabilities of features given each class were stored in logarithmic scale to mitigate underflow issues.

**Training Process**:

* For each class, the frequency of each feature was computed and smoothed using Laplace smoothing.
* Log probabilities for features conditioned on the class were calculated and stored for efficient inference computation.

**Prediction**:

* During prediction, the log of the joint probability was computed for each class, and the class with the highest probability was selected.

**Evaluation Metrics**:

* **Accuracy**: To measure overall correctness.
* **Recall**: To ensure relevant classes are not overlooked.
* **F1 Score**: To balance the measure in scenarios with class imbalance.

### **Conclusion**

This report highlights critical insights into user interactions, product placements, and click behavior based on a comprehensive dataset. Recommendations for future work include:

* Refining numerical feature preprocessing to enhance model performance.
* Exploring advanced techniques to address missing data.
* Conducting further analysis on gender-specific behavior for improved personalization.

### **Challenges and Future Directions**

**Challenges Faced**:

* Managing the sheer volume of data while ensuring computational efficiency.
* Addressing inconsistencies across data splits, particularly for PCA projections.
* Difficulty in performing a stratification on ‘click’ and ‘gender’ columns.

**Future Directions**:

* Extending the project to MLOps with a seamless pipeline.
* Developing a recommendation system for ad placement in impressions.
* Incorporating data warehousing for storing and analyzing the dataset.
* Facilitating business-level decision-making with a dynamic dashboard on Tableau or Power BI.

**References**:

1. Dataset description and preprocessing methods. Referenced from the FairJob dataset on Hugging Face (<https://huggingface.co/datasets/criteo/FairJob>).
2. Machine learning techniques applied to user behavior analysis.
3. PCA and cosine similarity in feature reduction.