**Analyzing user behavior in social media to predict the age group for e-commerce recommendations – a case study for Machine Learning Classification**

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| Sl NO | Topic | Page No |
| 1 | Overview | 3 |
| 2 | Business problem goals | 3 |
| 3 | Topic survey in depth | 6 |
| 4 | Critical assessment of topic survey | 7 |
| 5 | Methodology to be followed | 8 |
| 6 | References | 11 |

**Project Details**

# OVERVIEW

# Trell is a social media app that allows users to discover, create, and share short-form video content. The app was launched in India in 2020 and has since expanded to other countries. The platform is similar to other popular social media apps like TikTok and Instagram Reels, but with a focus on the Indian market.

# Trell's content is primarily centered around lifestyle and entertainment, with a variety of categories such as food, fashion, beauty, travel, and more. Users can create and share videos that showcase their interests and talents, and engage with other users by following, liking, and commenting on their content. The app also offers features like in-app purchases, challenges, and live streaming to keep users engaged and entertained.

# Analyzing User Behavior to Investigate Age Group Impact on Video Engagement" is an investigation into the relationship between user age and video engagement on a social media app. The project uses a dataset of user behavior on the social media app Trell, which includes data on user demographics, behavior, and engagement.

# Our goal is to develop a machine learning model to predict the age group of the Trell app users and use that data for further improvements in the app.

# BUSINESS PROBLEM STATEMENT (GOALS)

**BUSINESS PROBLEM UNDERSTANDING**

The goal of this project is to predict the age of users in a social media application and explore the effect of their age on their behavior and engagement with the platform. By understanding the age distribution of the user base, the social media company can tailor its products and services to meet the needs and preferences of different age groups. It can also be used for E-Commerce recommendation (digital marketing).

**BUSINESS OBJECTIVE**

The business objective of this project is to develop a predictive model using classification analysis that accurately estimates the age group of users based on their behavior and engagement on the social media platform. This model will enable the company to gain insights into the age distribution of their user base and the impact of age on user behavior.

Using this information, the social media company can design targeted marketing campaigns, develop new features and services that cater to different age groups, and improve the overall user experience. The business objective of this project is aligned with the company's vision of providing a personalized and engaging social media experience to its users.

Other potential business objectives for this project include identifying trends and patterns in user behavior across different age groups, improving the effectiveness of targeted advertising campaigns, and enhancing the company's ability to retain and attract users of all ages. Ultimately, the business objective should be designed to provide tangible benefits to the company and its users by improving user engagement and satisfaction, and increasing revenue and profitability.

**APPROACH**

Our Approach is to perform various methods on the Dataset which includes Understanding the Dataset, Data pre-processing, Handling Data Types, Scaling & Transformation of Data wherever required, dealing with the null values if required, Handling of Outliers with help of EDA techniques & Visualizations, Building Model, Evaluate & Deployment of the model.

**a. Descriptive Analysis:**

Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data. Measures of variability help communicate the spread of distribution by describing the shape and spread of the data set.

**b. Inferential Analysis:** Validating the inferences which are found with the help of descriptive analysis (Graphs) with the help of respective statistical tests if needed.

**c. Treating Outliers:**

Outlier is an observation in the data that lies at an abnormal distance from other values. Presence of an outlier may skew the results. Hence it is necessary to remove them. The interquartile range is a measure of outlier treatment. It is the difference between the third quartile and the first quartile. The IQR gives the range of middle 50% of the data which is free from outliers.

**d. Treating Missing Values:**

Detecting and handling missing values in the correct way is important, as they can impact the results of the analysis, and there are algorithms that can’t handle them. There are cases when a variable has a lot of missing values. In that case, we can drop the variable, if the variable is not a very important predictor for the target variable. For the categorical variable, the missing values can be replaced by the most frequent class of the variable. For the numeric variable, missing values can be replaced by the mean/ median.

**Standard missing values:**

The missing values which are detected by python are called standard missing values in python. The missing values detected by python include `Nan` and blank spaces. Non-Standard missing values: The missing values such as `? `, `-`, `NA` are not detected by python and are known as the non-standard missing values.

**e. Encoding Categorical Variables:**

Since, machine learning models are based on Mathematical equations and we can intuitively understand that it would cause some problem if we can either keep the Categorical data by encoding the categorical variable or we can drop by checking whether we need the variable for further modeling process because we would only want numbers in the equations. f. Dropping Unnecessary Columns: We are removing the columns which do not contribute to the model building or the columns which are of less, or of no importance.

**CONCLUSIONS:**

After the data preprocessing, a machine learning model was built using linear regression analysis to predict the age of users in a social media application and explore the effect of their age on their behavior and engagement with the platform.

This model will help the social media company gain insights into the age distribution of their user base and tailor their products and services to meet the needs and preferences of different age groups. It will also enable the company to design targeted marketing campaigns, develop new features and services, and improve the overall user experience.

In addition to this, we are using AI explain ability techniques to understand the reasons behind age-related differences in user behavior and engagement on the social media platform.

This will enable the company to make informed decisions about product development, marketing strategies, and user engagement efforts.

Overall, the predictive model and AI explain ability techniques developed in this project will help the social media company improve user engagement and satisfaction, increase revenue and profitability, and protect their reputation in a competitive market.

**TOPIC SURVEY IN BRIEF**

1. **PROBLEM UNDERSTANDING**

Understanding the age of users in a social media application can have significant implications for both the users and the developers. For users, their age can affect their online experience, as well as their privacy and security. For developers, understanding user age can inform the design of the application, as well as marketing and advertising efforts. Developing ML models on the dataset provides information about users and their age, along with various other features that can be used to build predictive models to identify which factors affect user age.

**2. CURRENT SOLUTION TO THE PROBLEM**

**a. Age verification**: Age verification is a crucial step for social media applications to ensure that underage users are not accessing the platform. This can be done through various methods, such as requiring users to provide their date of birth or uploading a government-issued ID for verification purposes.

**b. User engagement**: Social media applications often use engagement metrics, such as likes, comments, and shares, to understand user behavior and preferences. This data can provide insights into the age groups that are most active on the platform and inform future development efforts.

**3. PROPOSED SOLUTION TO THE PROBLEM**

The proposed solution to the problem is a linear regression-based machine learning model. It can be implemented by using various linear regression algorithms, such as simple linear regression, multiple linear regression, and polynomial regression. First, we are going to perform the data pre-processing step under which data cleaning, feature selection, feature scaling, and outlier detection steps are performed. Then, we will proceed to build a model that can predict the age of users based on their characteristics and behavior on the platform.

**4. POTENTIAL BENEFITS OF THE PROPOSED SOLUTION**

The potential benefits of the proposed solution are:

**Improved user experience**: By understanding user age and behavior, social media applications can tailor their platform to meet the needs and preferences of their user base, leading to a better user experience.

**Targeted advertising**: By identifying the age groups that are most active on the platform, developers can create targeted advertising campaigns that are more likely to be effective and generate revenue.

**Enhanced privacy and security**: By verifying user age, social media applications can protect underage users from inappropriate content and interactions, as well as protect user data from being accessed by unauthorized parties.

**Informed development decisions**: By using data to inform development decisions, social media applications can create new features and improvements that are more likely to be successful and meet user needs.

**5. REFERENCE TO THE PROBLEM**

https://www.kaggle.com/datasets/adityak80/trell-social-media-usage-data?select=train\_age\_dataset.csv

**CRITICAL ASSESSMENT OF TOPIC SURVEY**

Age group can have a significant impact on social media engagement. Different age groups may have different preferences, habits, and motivations when it comes to using social media platforms, which can affect how they engage with content and interact with other users. It is also important to note that age group impact on social media engagement can vary depending on the social media platform and the type of content. For example, younger age groups may be more likely to engage with video content on TikTok or Instagram Reels, while older age groups may be more likely to engage with news or educational content on Facebook or LinkedIn. Therefore, it is important to conduct platform-specific research to fully understand the impact of age groups on social media engagement.

Our project uses the user behavior on the social media app Trell. Based on this data, the user’s age group is predicted by using machine learning models. This prediction helps in improving the social media content, targeted marketing and building a recommendation system.

**METHODOLOGY TO BE FOLLOWED**

# 

**Business Understanding**:

It’s all about understanding the overview, the aspects of business activities & the necessary problems which the business is facing. To understand how to help the banks to identify its target customers and how to focus and change the method of approach for other customers to involve into term subscriptions.

**Data understanding**:

It involves study of data, shape, data types, number of rows and columns, type of columns and categories them into numerical and categorical data.

**Variables/Attributes in the data:**

The Trell social media usage dataset contains several variables (also known as features or columns) that provide information about user activity and behavior on the platform. The dataset includes the following variables:

userId : Unique number given to each user

tier : Tier of the city in which the user is residing

gender : Categorical feature representing the gender of the user. 1 represents male and 2 represents female

following\_rate : Number of accounts followed by the user(feature is normalized)

followers\_avg\_age : Average of age groups of all the followers of the user

following\_avg\_age : Average of age groups of all the accounts followed by the user

max\_repetitive\_punc : Maximum repetitive punctuations found in the bio and comments of the user

num\_of\_hashtags\_per\_action : Average number of hashtags used by the user per comment

emoji\_count\_per\_action : Average number of emojis used by the user per comment

punctuations\_per\_action : Average number of punctuations used by the user per comment

number\_of\_words\_per\_action : Average number of words used by the user per comment

avgCompletion : Average watch time completion rate of the videos

avgTimeSpent : Average time spent by the user on a video in seconds

avgDuration : Average duration of the videos that the user has watched till date

avgComments : Average number of comments per video watched

creations : Total number of videos uploaded by the user

content\_views : Total number of videos watched

num\_of\_comments : Total number of comments made by the user(normalized)

weekends\_trails\_watched\_per\_day : Number of videos watched on weekends per day

weekdays\_trails\_watched\_per\_day : Number of videos watched on weekdays per day

slot1\_trails\_watched\_per\_day : The day is divided into 4 slots. This feature represents the average number of videos watched in this particular time slot

slot2\_trails\_watched\_per\_day : The day is divided into 4 slots. This feature represents the average number of videos watched in this particular time slot

slot3\_trails\_watched\_per\_day : The day is divided into 4 slots. This feature represents the average number of videos watched in this particular time slot

slot4\_trails\_watched\_per\_day : The day is divided into 4 slots. This feature represents the average number of videos watched in this particular time slot

avgt2 : Average number of followers of all the accounts followed by the user

age\_group : This is a categorical feature denoting the age of the user. Age of users is divided into 4 groups, 1: \<18y; 2: 18-24y; 3: 24-30y; 4: >30y

These variables can be used to gain insights into user behavior and preferences on the Trell platform.

Based on the dataset's structure and available attributes, some possible methods that can be applied to analyze this dataset include:

**Data Cleaning** : The dataset may contain missing values, incorrect or inconsistent data, and outliers. Data cleaning techniques can be applied to remove such data points to ensure the quality of data used for analysis.

**Exploratory Data Analysis (EDA)** : EDA involves the analysis of the dataset to summarize its main characteristics and identify patterns, relationships, and trends. EDA techniques such as data visualization, summary statistics, and correlation analysis can be applied to explore the dataset.

**Sentiment Analysis** : Since the dataset contains user-generated content, sentiment analysis can be performed to identify the polarity of the text (positive, negative, or neutral). This can help in understanding user preferences, feedback, and opinions on different social media topics.

**Clustering** : Clustering algorithms can be applied to group users based on their social media usage patterns. This can help in identifying different user segments and understanding their behavior on the platform.

**Classification** : Classification models can be trained to predict user behavior, such as whether a user is likely to engage with a particular type of content or share it on their social media profile.

**REFERENCES**

1. [DOMAIN Presentation PPT](https://docs.google.com/presentation/d/1I77gXlQZuxKjd-8usmsQ7z90SLVhENU8AzsIAE6GMwM/edit?usp=sharing)
2. <https://www.ibm.com/in-en/topics/social-media-analytics>
3. <https://www.researchgate.net/publication/259148570_The_Power_of_Social_Media_Analytics>
4. <https://www.targetinternet.com/resources/how-different-age-groups-are-using-social-media>
5. [https://www.simplilearn.com/intersection-of-data-science-and-social-media-article#:~:text=Social%20media%20analytics%20and%20bigcontent%20based%20on%20customer%20sentiment](https://www.simplilearn.com/intersection-of-data-science-and-social-media-article#:~:text=Social%20media%20analytics%20and%20big,content%20based%20on%20customer%20sentiment).
6. [Data Science for Humanity](https://datascience.stanford.edu/research/research-areas/data-science-humanity#:~:text=Data%20science%20methods%20and%20approaches,%2C%20countries%2C%20and%20societies%3B%20from)
7. http://jenpan.com/jen\_pan/50c.pdf

**Notes For Project Team**

|  |  |
| --- | --- |
| Original owner of data | Trell |
| Data set information | The dataset was posted in a hackathon sponsored by Trell themselves in 2022. The dataset contains various features of more than 4 lakh users. Some prominent features include factors like age, time spent, gender, count of words etc.. Using these features we predict the age group of a user. |
| Any past relevant articles using the dataset | <https://www.kaggle.com/datasets/adityak80/trell-social-media-usage-data?select=train_age_dataset.csv> |
| Reference | [AI-planet](https://aiplanet.com/challenges/46/cascade-cup-data-science-hackathon/data) |
| Link to web page | <https://aiplanet.com/challenges/46/cascade-cup-data-science-hackathon/data> |

### 

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*