Bangalore Institute of Technology

Department of Computer Science and Engineering

K R Road, V V Puram, Bengaluru-560004

Mini Project Synopsis on

Social Network Ad Classification

Submitted as the mini project for the subject
Data Mining and Data Warehousing (18CS641)

Submitted by

Samarth S H	**1BI20CS150**
Vamshik R	**1BI20CS184**
Rakshith Rajesh N B	**1BI21CS408**

For academic year 2022-23

Under the guidance of
Dr. Bhargavi M S
Associate Professor

INTRODUCTION:

We know that the businesses in today's world runs on strategized marketing and advertising. The goal of all commercial businesses is to expand their customer base and reach more buyers. The main task of the Social Network Ad classification is to sort these advertisements into different classes based on various factors like their content, type and usability of product/service, etc. The crucial outcome of this analysis is to accurately classify the ads and the targeted audiences it's reaching which leads to the customer buying or not buying the product.

Classifying social media ads to find the most profitable customers is a necessary strategy these days. Sometimes, the product you are offering is not suitable for all people when it comes to age and income. For example, a person between the ages of 20 and 25 may like to spend more on smartphone covers than a person between the ages of 40 and 45.

The main objectives of this analysis are:

- Forecast the sales of the product and adjust to the demand levels
- Predict the profits and achieve the rates
- To develop enhanced customer targeting strategies
- Study the purchase pattern of the customers of different classes
- To follow a trend/change in the market

The algorithms of classification and prediction models help us determine whether the customer is likely to buy the product or not. This data can be used to find new ways to grab the customer's attention. The classification algorithms also help us identify the emerging trends in the buying pattern of the customers of a particular class. Such a classification model can help a business in ways beyond the present scope of the market.

The social network ad classification problem plays a crucial role in ensuring both the user satisfaction.

Dataset Description

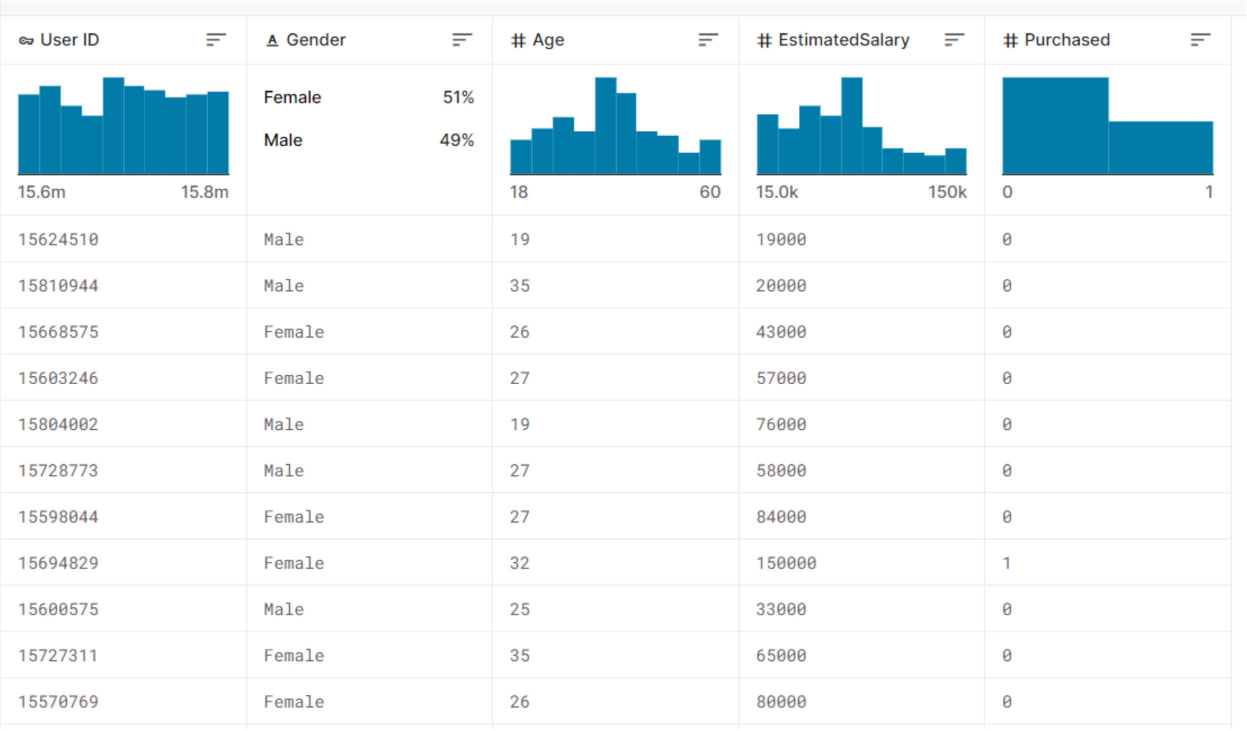
Dataset Selection: We will carefully select diverse datasets from various domains, ensuring they represent real-world scenarios and challenges. These datasets will encompass different dimensions and characteristics, allowing us to evaluate the effectiveness of classifier across various contexts.

Dataset consists of:

- 400 entries of non-null data
- 5 data columns: User ID, Gender, Age, Estimated Salary, Purchased
- Purchased column is of the binary type, where 0 indicated not purchased and 1 indicates purchased
- Gender is categorical type having two values –Male and Female

- A total of 257 “not purchased” and 143 “purchased” entries are in the dataset
- Age ranges from a minimum of 18 years to a maximum of 60 years
- Dataset is split into two parts: one for training and one for testing in the ratio of 75% and 25%.

Snapshots of Dataset



User ID	Gender	# Age	# EstimatedSalary	# Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	0
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	0
15728773	Male	27	58000	0
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	0
15727311	Female	35	65000	0
15570769	Female	26	80000	0

Preprocessing Techniques to be used

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. Some of the most efficient pre-processing techniques used in the insurance claim prediction is:

- **Removal of any Duplicate rows (if any):**

  In machine learning, it is important to distinguish the matrix of features (independent variables) and dependent variables from dataset.

- **Check for empty elements:**

The next step of data pre-processing is to handle missing data in the datasets. If our dataset contains some missing data, then it may create a huge problem for our machine learning model. Hence it is necessary to handle missing values present in the dataset.

There are mainly two ways to handle missing data, which are:

1. By deleting the particular row: The first way is used to commonly deal with null values. In this way, we just delete the specific row or column which consists of null values. But this way is not so efficient and removing data may lead to loss of information which will not give the accurate output.
2. By calculating the mean: In this way, we will calculate the mean of that column or row which contains any missing value and will put it on the place of missing value. This strategy is useful for the features which have numeric data such as age, salary, year, etc. Here, we will use this approach.

- To handle missing values, we will use Scikit-learn library in our code, which contains various libraries for building machine learning models. Here we will use Imputer class of sklearn.preprocessing library.
- **Converting categorical Columns to Numeric columns using One-Hot Binary Encoding:**
  - One hot encoding is a technique that we use to represent categorical variables as numerical values in a machine learning model. The input to this transformer should be an array-like of integers or strings, denoting the values taken on by categorical (discrete) features. The features are encoded using a one-hot (aka 'one- of-K' or 'dummy') encoding scheme. This creates a binary column for each category and returns a sparse matrix or dense array (depending on the sparse_output parameter)
  - By default, the encoder derives the categories based on the unique values in each feature. Alternatively, you can also specify the categories manually.
  - This encoding is needed for feeding categorical data to many scikit-learn estimators, notably linear models and SVMs with the standard kernels.
- **Scaling data to the same range**:
  Scaling data to the same range means adjusting the values of different variables or attributes so that they have a common scale, such as 0–1 or -1–1 .
- **Reducing dimensionality**

  As the number of dimensions increases, the more difficult it is to find strict differences between instances. This phenomenon is known as the curse of dimensionality.

Classification Techniques to be applied

1. **Decision Trees (DTs)**
   - They are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

- There are two nodes, which are the **Decision Node** and **Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

- It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
- Algorithm of the decision tree is as follows:
  - i. Begin the tree with the root node, says S, which contains the complete dataset.
  - ii. Find the best attribute in the dataset using **Attribute Selection Measure (ASM)**
  - iii. Divide the S into subsets that contains possible values for the best attributes
  - iv. Generate the decision tree node, which contains the best attribute.
  - v. Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.
- **Attribute Selection Measure**
  - ➢ There are two techniques to select the best attribute:
    - **i. Information Gain:** measurement of changes in entropy after the segmentation of a dataset based on an attribute.
    - **ii. Gini Index:** It is a measure of impurity or purity used while creating a decision tree

2. **K-Nearest Neighbors**
   - It is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
   - It assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
   - stores all the available data and classifies a new data point based on the similarity
   - It also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
   - The working of the KNN algorithm is as follows
     - **i.** Select the number K of the neighbors
     - **ii.** Calculate the Euclidean distance of **K number of neighbors**
     - **iii.** Take the K nearest neighbors as per the calculated Euclidean distance
     - **iv.** Among these k neighbors, count the number of the data points in each category.
     - **v.** Assign the new data points to that category for which the number of neighbor is maximum.
   - **Selecting a value for K:** There is no particular way to determine the best value for "K", so we need to try some values to find the best out of them. The most preferred value for K is 5.

Tools to be used

- Jupyter Notebook
- Kaggle
- MatPlotlib and Seaborn