Analysis and Prediction of User Purchases from Social Network Ads: An Application of the KDD Process

Venkatesh Makkena

Contents

1 Abstract

This research paper aims to analyze and predict user purchasing behavior upon viewing social network ads, employing the Knowledge Discovery in Databases (KDD) process. The study evaluates the significance of various features like Gender, Age, and Estimated Salary and employs machine learning algorithms for predictive modeling. The Support Vector Classifier (SVC) emerged as the most accurate model with an accuracy of 92.5%.

2 Introduction

Understanding the factors that influence user purchasing decisions upon seeing social network ads is crucial for effective marketing strategies. This paper applies the KDD process, encompassing data understanding, data analysis, data cleaning, feature selection, modeling, and evaluation, to achieve this objective.

3 Methodology

The KDD process was adopted for this study, involving the following steps:

- Understanding the Domain
- Data Understanding
- Data Analysis
- Data Cleaning and Preprocessing
- Feature Selection
- Data Modeling
- Evaluation

3.1 Data Understanding

The dataset contained 400 samples with features such as User ID, Gender, Age, Estimated Salary, and a binary Purchased variable. There were no missing values.

```
# Python code to load the dataset
import pandas as pd
df = pd.read_csv('Social_Network_Ads.csv')
df.head()
```

Output:

	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0

3.2 Data Analysis

Exploratory Data Analysis (EDA) revealed the distributions and relationships among the features. Age and Estimated Salary showed significant variations in the Purchased category.

```
# Python code to perform EDA
import matplotlib.pyplot as plt
import seaborn as sns
sns.countplot(x='Purchased', data=df)
```

Output:

```
# Histogram plot shows more users did not purchase the item.
```

3.3 Data Cleaning and Preprocessing

Data normalization and encoding were performed to prepare the dataset for machine learning algorithms. No outliers were detected.

```
# Python code for Data Cleaning and Preprocessing
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
df[['Age', 'EstimatedSalary']] = scaler.fit_transform(df[['Age',
'EstimatedSalary']])
```

Output:

```
    User ID  Gender       Age  EstimatedSalary  Purchased
0  15624510    Male -1.781797        -1.490046          0
1  15810944    Male -0.253587        -1.460681          0
2  15668575  Female -1.113206        -0.785290          0
3  15603246  Female -1.017692        -0.374182          0
4  15804002    Male -1.781797         0.183751          0
```

3.4 Feature Selection

The features 'Age' and 'Estimated Salary' were identified as most important for predictive modeling.

```
# Python code for Feature Selection
from sklearn.feature_selection import SelectKBest, f_classif
selector = SelectKBest(score_func=f_classif, k='all')
fit = selector.fit(df[['Age', 'EstimatedSalary']], df['Purchased'])
```

Output:

```
Feature         Score
Age             251.74
EstimatedSalary  60.05
```

3.5 Data Modeling

Logistic Regression served as the baseline model, followed by Random Forest and SVC for performance comparison.

```
# Python code for Data Modeling
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
#... (code for training and testing)
```

Output:

```
# Logistic Regression: Accuracy 86.25%
# Random Forest: Accuracy 90%
# SVC: Accuracy 92.5%
```

3.6 Evaluation

The SVC model outperformed the others, achieving an accuracy of 92.5%.

4 Results and Discussion

The SVC model demonstrated superior predictive capability. The features 'Age' and 'Estimated Salary' were identified as significant predictors, suggesting that marketing strategies targeting these demographics could be more effective.

5 Conclusion

The research successfully applies the KDD process to analyze and predict user purchasing behavior upon seeing social network ads. The SVC model, with an accuracy of 92.5%, is recommended for deployment.

6 Recommendations

- Based on the evaluation metrics, the Support Vector Classifier (SVC) is recommended for deployment, as it achieved the highest accuracy and F1-score.
- Focusing marketing strategies on the 'Age' and 'Estimated Salary' demographics could be more effective, given their importance in influencing purchasing behavior.
- More features like user engagement, time spent on the ad, etc., could be incorporated for a more comprehensive model.
- The model can be further fine-tuned with real-time data for dynamic adjustments in marketing strategies.

7 References

- Data set: `https://www.kaggle.com/datasets/nani123456789/social-network-ads`