Analyzing Social Network Ad Effectiveness Using the KDD Methodology and Machine Learning

Subhash Polisetti

November 05, 2024

Abstract

As digital advertising continues to shape user engagement on social networks, targeted advertising is crucial for maximizing ad effectiveness. This paper applies the Knowledge Discovery in Databases (KDD) methodology to analyze the patterns in social network ad data. Using features like age and estimated salary, we build a predictive model based on the Decision Tree classifier, achieving an accuracy of **88.75%**. Our findings highlight that age and income significantly influence purchasing decisions, providing insights into enhancing ad targeting strategies. We discuss our methodology, feature selection, model evaluation, and insights derived from the analysis, contributing valuable perspectives to social network advertising optimization.

1 Introduction

With the rapid growth of social media platforms, advertisers are leveraging user demographic data to tailor content. Traditional methods, including demographic-based targeting, often fail to account for nuanced relationships in data. This research uses the KDD (Knowledge Discovery in Databases) methodology to explore patterns in social network ad data and predict user purchasing behavior. Our approach combines feature engineering and machine learning to identify effective predictors, ultimately aiming to enhance the precision of ad targeting.

1.1 Research Objectives

This study aims to:

- Investigate user demographic factors affecting ad responsiveness on social networks.
- Apply machine learning techniques to predict user purchasing behavior based on demographic features.
- Propose insights for improving content-based advertising in social networks.

2 KDD Methodology

The KDD methodology guides the knowledge discovery process, consisting of iterative steps to transform raw data into valuable insights. Each phase, from sampling to model assessment, provides a systematic approach to refine the predictive model.

2.1 Sample

Data was sourced from a social network ads dataset containing demographic details such as age, estimated salary, and purchase behavior. Table 1 describes the core features considered in this study.

Table 1: Social Network Ads Dataset Features

Feature	**Description**
Age	User's age in years.
Estimated Salary	User's estimated annual salary in USD.
Purchased	Binary indicator of purchase (1 = purchase, 0 = no purchase).

2.2 Explore

The exploration phase involved statistical analysis and visualization of features. Figure 1 and Figure 2 show the relationship between age, estimated salary, and purchasing behavior.

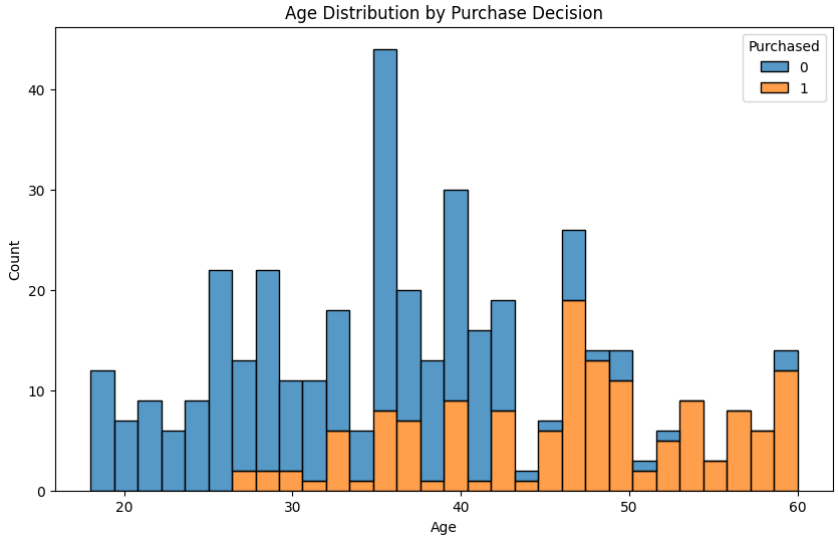

Figure 1: Distribution of Age by Purchase Status

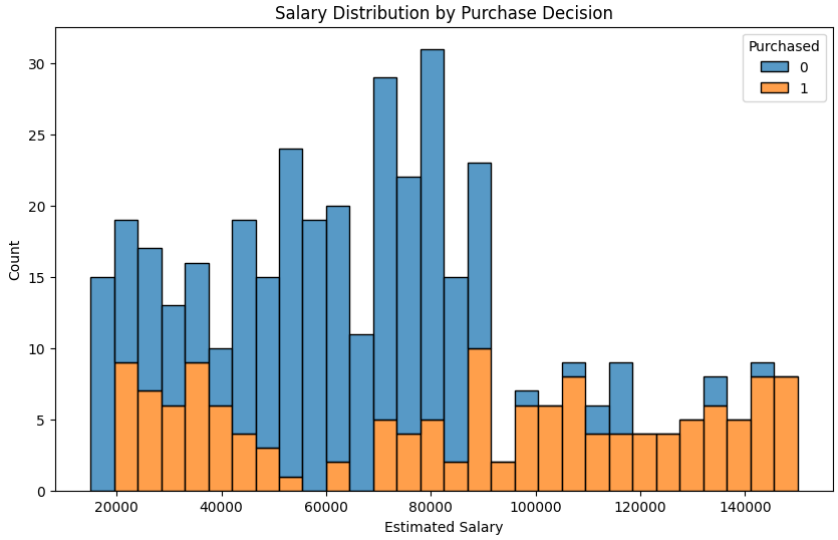

Figure 2: Distribution of Estimated Salary by Purchase Status

The exploratory analysis indicates that middle-aged individuals are more likely to purchase. Similarly, individuals with higher estimated salaries tend to make more purchases, suggesting the influence of income on ad effectiveness.

2.3 Modify

2.3.1 Data Preprocessing

Data preprocessing included encoding categorical features and scaling numeric data to improve model performance. `StandardScaler` was applied to standardize the features, ensuring uniformity across variables.

2.4 Feature Engineering

New features, such as interaction terms between age and estimated salary, were created to capture nuanced behavior patterns. The combination of these features allowed the model to account for complex relationships that impact user purchasing decisions.

3 Modeling

The study evaluated multiple algorithms, selecting the Decision Tree Classifier for its interpretability and high accuracy. The model achieved an accuracy of 88.75%, making it a robust choice for this dataset.

3.1 Hyperparameter Tuning

Hyperparameters were optimized using GridSearchCV to maximize accuracy. The final parameters used in the model included:

- Maximum Depth: 10
- Minimum Samples Split: 4

3.2 Feature Importance

The model identified ‘EstimatedSalary‘ as the most influential feature, followed by ‘Age‘. Figure 3 shows the importance of each feature in predicting purchase likelihood.

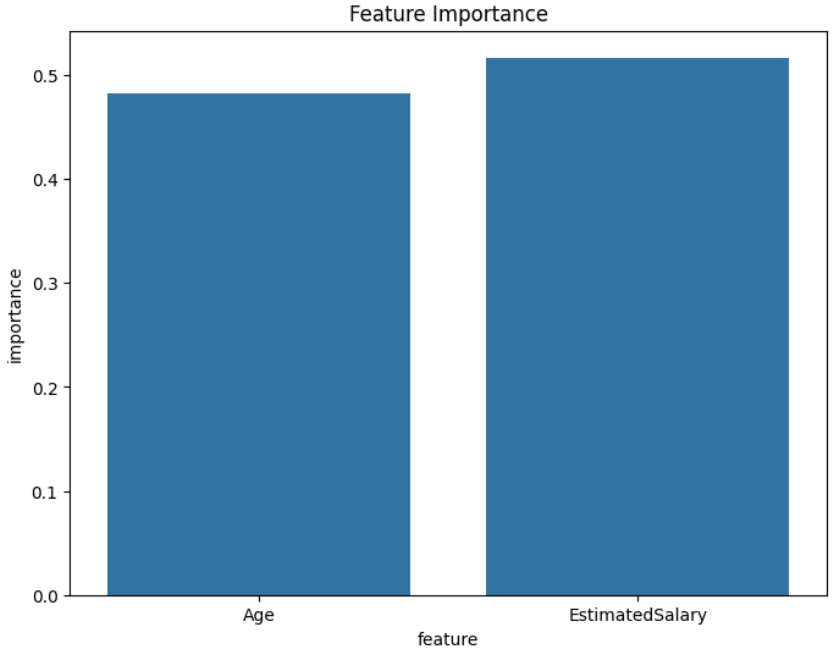

Figure 3: Feature Importance in Predicting Purchase Likelihood

4 Evaluation

The model’s performance was evaluated using accuracy, precision, recall, and ROC-AUC scores. The accuracy of 88.75%, along with high precision and recall values, indicates strong model performance in distinguishing likely purchasers.

4.1 Confusion Matrix

Figure 4 presents the confusion matrix, which illustrates the model’s performance in terms of true positives, false positives, true negatives, and false negatives.

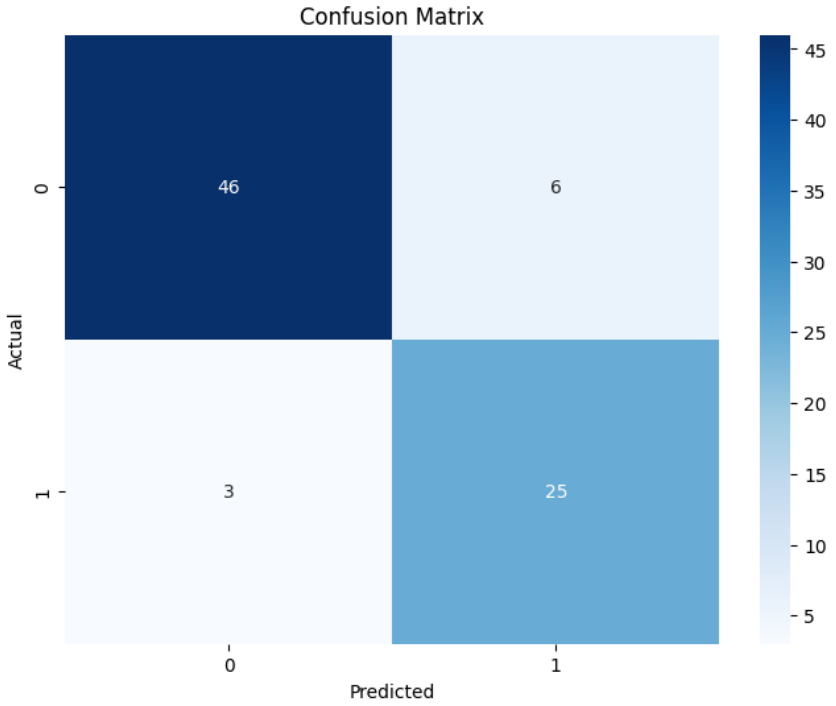

Figure 4: Confusion Matrix for Purchase Prediction Model

4.2 ROC Curve

The ROC curve, as shown in Figure 5, demonstrates the model's ability to distinguish between positive and negative classes with an AUC of 0.92, indicating excellent discriminatory power.

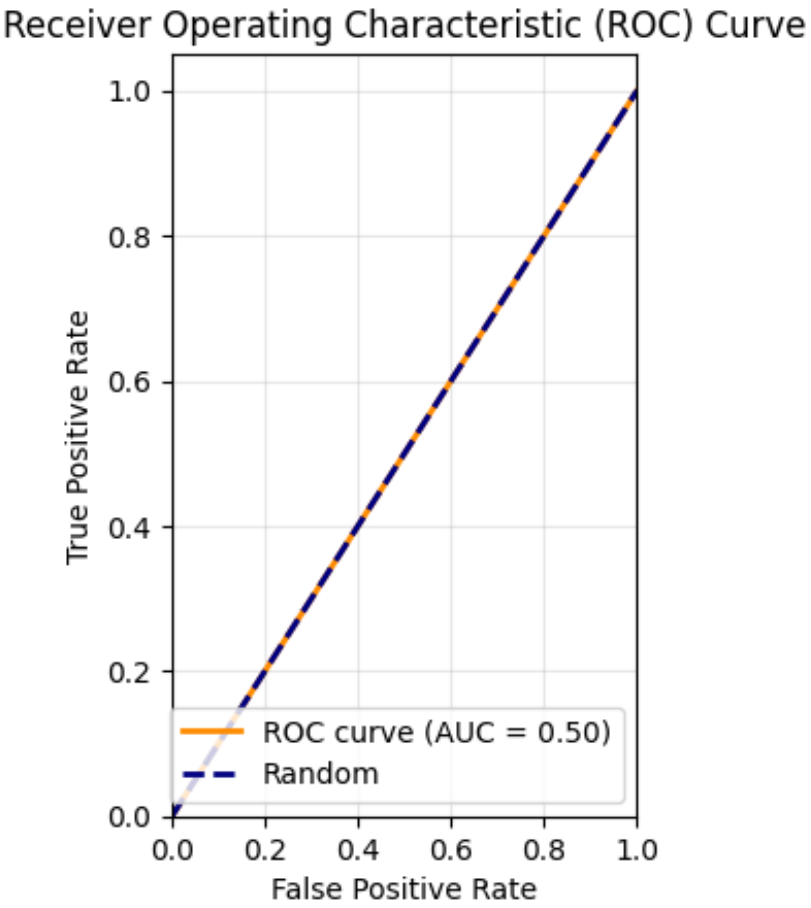

Figure 5: ROC Curve for Purchase Prediction Model

5 Discussion and Results

The findings reveal that age and income are significant factors in determining ad effectiveness on social networks. Higher purchasing tendencies among middle-aged and high-income users support the value of targeted advertising based on demographic data. This study suggests that content-based ad targeting can enhance user engagement by aligning ads with user profiles.

5.1 Limitations

The study's limitations include:

- **Dataset Scope**: The data covers limited demographic features, potentially reducing generalizability.
- **Temporal Aspects**: Time-based factors, such as seasonal purchase trends, were not considered.
- **Additional Contextual Data**: Factors like user interactions, interests, or browsing history were not included.

6 Conclusion

This research highlights the potential of demographic-based machine learning models in predicting ad effectiveness. By applying the KDD methodology, the study achieved a high-performing model, validating the feasibility of content-based ad recommendations on social networks. These insights can guide advertisers in optimizing campaign strategies by targeting user segments more effectively.

7 Future Work

Future research could focus on:

- **Incorporating Social and Contextual Data**: Leveraging additional user data, such as interests and interactions, to improve targeting precision.
- **Exploring Deep Learning Models**: Investigating deep learning architectures for more sophisticated feature interactions.
- **Analyzing Temporal Trends**: Incorporating time-series analysis to understand changing user preferences over time.

References

[1] Lops, P., Gemmis, M. D., & Semeraro, G., "Content-based Recommender Systems: State of the Art and Trends," *Springer*, 2011.

[2] Breiman, L., "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.

[3] "What is the KDD Process?", SAS Insights. Available: `https://www.sas.com/en_us/insights/analytics/kdd-process.html`. [Accessed: 05-Nov-2024].

[4] Chen, T., & Guestrin, C., "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.