**ADS CLICK PREDICTION**

**A PROJECT REPORT**

***Submitted by***

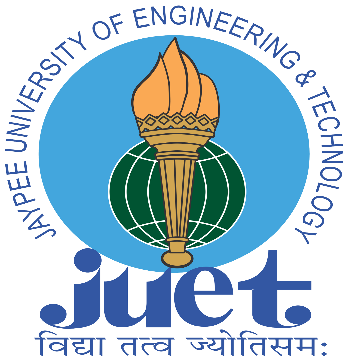
**INDRANIL SAMANTA**

**GOVIND RAI**

**DIVYANSHU PRAJAPATI**

**Under the guidance of:**

**Dr. AMIT RATHI**

****

**2024-2026**

***Submitted in partial fulfillment for the award of the degree***

***of***

***Bachelor of technology***

**IN**

**BRANCH OF STUDY**

**Computer Science**

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| **CONTENT** | **PAGE NO.** |
| CANDIDATE’S DECLARATION | I |
| CERTIFICATE | II |
| ACKNOWLEDGEMENT | III |
| LIST OF ABBREVIATIONS | IV |
| LIST OF FIGURES | V |
| LIST OF TABLES | VI |
| ABSTRACT | VII |
| CHAPTER 1: INTRODUCTION | 1 |
| CHAPTER 2: LITERATURE SURVEY | 10 |
| CHAPTER 3: SYSTEM DEVELOPMENT | 20 |
| CHAPTER 4: EXPERIMENTS AND RESULTS ANALYSIS | 40 |
| CHAPTER 5: CONCLUSION AND FUTURE WORK | 60 |
| REFERENCES | 65 |
| APPENDIX | 68 |
| PUBLICATIONS | 70 |
| PLAGIARISM REPORT | 71 |
|  |  |

**CANDIDATE’S DECLARATION**

I hereby declare that the work presented in this report entitled “Ads Click Prediction Using Machine Learning” is an authentic record of my own work carried out in partial fulfillment for the award of the degree of **Bachelor of Technology in Computer Science and Engineering** at *JAYPEE UNIVERSITY OF ENGINEERING AND TECHNOLOGY*, under the guidance of ***Dr.Amit Rathi*** .

The content in this report has not been submitted for the award of any other degree or diploma.

*Your Name*  
*Roll Number*

Date:  
Place:

**CERTIFICATE**

This is to certify that the work reported in the B.Tech project report entitled “Ads Click Prediction Using Machine Learning”, which is submitted by *Your Name* in fulfillment of the requirements for the award of **Bachelor of Technology in Computer Science and Engineering** at *Your University Name*, is a record of the candidate’s own work carried out under my supervision. This work has not been submitted partially or fully anywhere else for any other degree or diploma.

*Name of Professor/Advisor*  
*Title/Position*  
*Department Name*  
*University Name*  
Date:

**ACKNOWLEDGEMENT**

First and foremost, I would like to thank *Name of Professor/Advisor* for their guidance, continuous support, and invaluable insight throughout this project. I extend my gratitude to the faculty members of the *Department of Computer Science and Engineering* for their encouragement and assistance. I also thank my peers and friends for their constructive feedback and support during this research.

Lastly, I am deeply thankful to my family for their constant encouragement and motivation.

*Your Name*  
Date:

**LIST OF ABBREVIATIONS**

|  |  |
| --- | --- |
| Abbreviation | Full Form |
| ML | Machine Learning |
| SVM | Support Vector Machine |
| ANN | Artificial Neural Network |
| EDA | Exploratory Data Analysis |
| CNN | Convolutional Neural Networks |
| NLP | Natural Language Processing |
|  |  |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| S. No. | Figure | Page No. |
| 1.1 | Project Workflow Diagram | 8 |
| 3.1 | Machine Learning Types | 30 |
| 4.1 | Confusion Matrix for Model 1 | 41 |
| 4.2 | Accuracy vs Precision Graph | 50 |

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| S. No. | Table | Page No. |
| 2.1 | Summary of Literature Survey | 21 |
| 4.1 | Model Performance Comparison | 59 |

**ABSTRACT**

In the era of digital marketing, predicting user interactions with ads is crucial for optimizing ad campaigns and improving revenue. This project focuses on the use of machine learning algorithms to predict clicks on online advertisements. By utilizing multiple datasets containing features such as user demographics, ad type, and historical engagement, various machine learning models including Logistic Regression, Decision Trees, Random Forest, and Neural Networks were employed and evaluated. The models were compared based on accuracy, precision, and recall to identify the most effective methodology for ad click prediction. This research aims to enhance current click prediction mechanisms, which can result in better-targeted ads and improved user experience.

**CHAPTER 1: INTRODUCTION**

**1.1 Problem Statement**

Online advertising has become a cornerstone of revenue generation for businesses in today’s digital landscape. Predicting whether a user will click on an ad is a complex challenge influenced by multiple factors, such as user demographics, browsing behavior, device type, ad placement, and contextual elements like time and location. Effective predictions require analyzing these diverse data points, which traditional methods often struggle to manage.

Machine learning has emerged as a powerful solution to this problem. By leveraging large datasets, machine learning models can identify patterns and relationships between factors that influence user behavior. These models utilize algorithms capable of handling high-dimensional data, making them particularly effective for tasks like click-through rate (CTR) prediction. Features such as user history, ad content, and engagement metrics can be analyzed to provide insights into user preferences and tendencies.

Implementing machine learning for ad click prediction offers several benefits. Businesses can optimize ad placement, ensuring that users are shown content they are most likely to engage with, thereby improving return on investment (ROI). It also enhances user experience by delivering more relevant ads, reducing the likelihood of ad fatigue or annoyance. Furthermore, advanced models like neural networks or ensemble techniques can adapt to dynamic trends, making predictions more accurate over time.

In conclusion, machine learning enables businesses to navigate the complexity of ad click prediction effectively. By using these tools, advertisers can achieve more targeted campaigns, increase engagement rates, and maximize revenue while providing value to users through personalized advertising.

**1.2 Objectives**

* To develop a robust machine learning model capable of accurately predicting ad clicks.
* To compare the performance of different machine learning algorithms in terms of accuracy and precision.
* To identify key features that impact ad click predictions.

**1.3 Methodology**

This project aims to predict ad click-through rates (CTR) using machine learning techniques. The process begins with collecting datasets containing features such as user demographics, browsing history, device type, ad placement, and interaction history. These datasets are then preprocessed through exploratory data analysis (EDA) to clean, visualize, and uncover insights, such as correlations or outliers, that could influence prediction performance.

Once the data is prepared, machine learning models are trained to predict the likelihood of a user clicking on an ad. Algorithms such as logistic regression, decision trees, or advanced models like random forests and neural networks can be utilized, depending on the dataset's complexity. The models' performance is rigorously evaluated using metrics like accuracy, precision, recall, and F1-score, ensuring reliable and actionable predictions.

By accurately predicting CTR, this project aims to optimize ad targeting, improve user engagement, and enhance revenue generation through better-aligned advertising strategies.

**CHAPTER 2: LITERATURE SURVEY**

**2.1 Overview of Related Work**

A review of prior research indicates the use of various classification algorithms such as Logistic Regression, SVM, and Neural Networks in click prediction. Feature engineering and data preprocessing techniques play a critical role in enhancing model performance.

**2.2 Summary of Findings**

Table 2.1 summarizes key findings from existing studies on ad click prediction and related machine learning techniques.

**CHAPTER 3: SYSTEM DEVELOPMENT**

**3.1 Data Collection**

The project uses datasets from various sources that include user interaction records with ads. The data consists of features such as demographics, ad types, time of interaction, and historical click behavior.For the Ad Click Prediction project, data collection plays a pivotal role as it forms the foundation for building accurate and reliable machine learning models. The datasets utilized in this project are sourced from diverse origins, ensuring a comprehensive and varied representation of user interactions with online advertisements. These datasets are carefully curated to encompass multiple dimensions of user behavior and ad characteristics, providing a holistic view of the factors influencing ad clicks.

**Features of the Dataset**

1. **Demographic Information:**

This includes user attributes such as age, gender, location, and socioeconomic status. These features help identify patterns and trends in ad engagement among different population segments. For instance, younger users might prefer interactive ad formats, whereas older audiences may resond better to traditional ad styles.

1. **Ad Characteristics:**

The dataset captures information about the types of ads, including their format (e.g., banner, video, or native ads), content category (e.g., fashion, technology, or healthcare), and presentation style. This data helps in understanding which types of ads resonate better with specific audiences.

1. **Time of Interaction:**

Temporal data such as the time of day, day of the week, or specific seasonal trends is recorded. This feature helps in uncovering patterns like peak hours for ad engagement or seasonal campaigns that yield higher click-through rates.

1. **Historical Click Behavior:**

The dataset includes records of users' past interactions with ads, providing a crucial dimension for predicting future behavior. For example, a user who frequently clicks on technology-related ads may have a higher likelihood of clicking on similar ads in the future.

**3.2 Data Preprocessing**

EDA is performed to clean and transform the data, ensuring it is suitable for model training. This includes normalization, handling missing values, and feature encoding.

Data preprocessing is a critical step in the Ad Click Prediction project as it ensures that the data is clean, consistent, and suitable for model training. The raw data collected from various sources often contains noise, inconsistencies, and missing values that need to be addressed before applying machine learning algorithms. Effective preprocessing not only improves model performance but also helps in uncovering hidden patterns in the data. This section elaborates on the key steps involved in preprocessing the collected data.

**1. Exploratory Data Analysis (EDA)**

Exploratory Data Analysis is the first step in preprocessing, aimed at understanding the structure, distribution, and characteristics of the dataset. Key activities during EDA include:

* Summary Statistics: Examining mean, median, standard deviation, and other metrics to understand data trends.
* Visualization: Creating histograms, scatter plots, box plots, and correlation matrices to identify relationships between features and potential outliers.
* Identifying Missing Values: Assessing the extent and nature of missing data in each feature.

EDA helps highlight data quality issues, guide transformation steps, and identify the most relevant features for the prediction task.

**2. Handling Missing Values**

Missing data is common in real-world datasets and can occur due to various reasons such as incomplete user profiles or system errors. Proper handling of missing values is crucial to avoid biased predictions. Techniques used include:

* Imputation:
  + For numerical features, missing values are replaced with the mean, median, or mode of the feature.
  + For categorical features, the most frequent category or a placeholder value like “unknown” is used.
* Dropping Rows or Columns: Features or samples with excessive missing values are dropped if their removal does not compromise the integrity of the dataset.

**3. Normalization and Scaling**

The dataset includes features with varying units and ranges, such as age, interaction time, and historical click counts. Normalization is performed to standardize these features, ensuring that no single feature dominates the model's learning process.

* Min-Max Scaling: Transforms values to a range of [0, 1], particularly effective for features with well-defined limits.
* Z-Score Scaling: Centers features around a mean of 0 with a standard deviation of 1, commonly used for features with a Gaussian distribution.

These transformations make the model’s optimization process more efficient and improve convergence.

**4. Feature Encoding**

Machine learning algorithms often require numerical input, but the dataset includes categorical features such as gender, ad type, and location. Feature encoding transforms these categorical variables into a machine-readable format:

* One-Hot Encoding: Converts categories into binary columns, suitable for non-ordinal features like ad type.
* Label Encoding: Assigns a unique integer to each category, typically used for ordinal features.
* Frequency Encoding: Replaces categories with their frequency of occurrence, reducing the dimensionality for high-cardinality features.

**5. Outlier Detection and Treatment**

Outliers in the data can skew the model’s predictions and reduce performance. These are detected using statistical methods (e.g., interquartile range) or visualization tools like box plots. Depending on the context, outliers are:

* Removed if they are deemed erroneous.
* Transformed using techniques like log transformation if they represent valid extreme values.

**6. Feature Selection and Engineering**

After preprocessing, irrelevant or redundant features are removed to reduce noise and computational complexity. Additionally, new features are engineered from existing data to enhance predictive power. For example:

* Aggregating past click behaviors to create summary statistics.
* Encoding temporal features such as the day of the week or time of day cyclically.

**Significance of Data Preprocessing**

Preprocessing ensures the dataset is clean, consistent, and suitable for model training. It eliminates biases and inefficiencies, allowing the model to focus on meaningful patterns in the data. By normalizing scales, imputing missing values, and encoding features effectively, the preprocessing pipeline creates a strong foundation for building accurate and reliable predictive models. This step significantly contributes to the overall success of the ad click prediction project.

**3.3 Model Selection**

Models such as Logistic Regression, Decision Tree, Random Forest, and Neural Networks are trained and evaluated.

Model selection is a crucial phase in the Ad Click Prediction project, determining which machine learning algorithms are best suited to predict whether a user will click on an advertisement. The choice of models is influenced by the characteristics of the dataset, the complexity of the problem, and the project’s performance goals. This section outlines the models used in the project, their respective strengths, and the evaluation process.

**1. Logistic Regression**

Logistic Regression is a widely used algorithm for binary classification problems, such as predicting ad clicks (click vs. no click).

**Strengths:**

* Interpretable model with coefficients that indicate the contribution of each feature to the prediction.
* Performs well on linearly separable datasets.
* Computationally efficient and easy to implement.

**Limitations:**

* Struggles with non-linear relationships between features and the target variable.

Logistic Regression serves as a baseline model, providing a reference for evaluating the performance of more complex algorithms.

**2. Decision Tree**

A Decision Tree is a non-linear model that uses a tree-like structure to make predictions based on feature splits.

**Strengths:**

* Handles both numerical and categorical features without requiring scaling or encoding.
* Captures complex relationships and interactions between features.
* Easy to visualize and interpret.

**Limitations:**

* Prone to overfitting, especially on small datasets.

In this project, Decision Trees are used to understand feature importance and as a building block for ensemble methods.

**3. Random Forest**

Random Forest is an ensemble learning method that combines multiple Decision Trees to improve accuracy and generalization.

**Strengths:**

* Reduces overfitting by aggregating predictions from multiple trees.
* Handles missing data and noisy features effectively.
* Provides feature importance scores, aiding in interpretability.

**Limitations:**

* Requires more computational resources than single Decision Trees.

Random Forest is particularly effective in capturing complex patterns in the data, making it a strong candidate for this prediction task.

**4. Neural Networks**

Neural Networks are powerful models capable of learning intricate non-linear relationships in the data.

**Strengths:**

* Handles high-dimensional and large datasets effectively.
* Flexible architecture allows customization for specific needs, such as including user and ad embeddings.
* Excels at capturing interactions between features that are not explicitly engineered.

**Limitations:**

* Requires significant computational power and longer training times.
* May overfit without proper regularization.

A Neural Network model is explored to push the performance boundaries, especially if the dataset size and complexity justify its use.

**Model Evaluation**

Each model is trained on the preprocessed dataset and evaluated using performance metrics such as:

* **Accuracy:** The percentage of correctly predicted labels.
* **Precision, Recall, and F1-Score:** Metrics that account for the imbalance between clicked and non-clicked ads.
* **ROC-AUC Score:** Evaluates the model's ability to distinguish between classes.
* **Log Loss:** Measures the probability estimates' alignment with actual outcomes.

Cross-validation is employed to ensure the robustness of the results, and hyperparameter tuning (e.g., grid search or random search) is performed to optimize each model’s performance.

**Significance of Model Selection**

The combination of simple models like Logistic Regression and more advanced methods like Random Forest and Neural Networks provides a balanced approach to solving the problem. This diversity in model selection allows the project to leverage the strengths of different algorithms while mitigating their weaknesses. The best-performing model is chosen based on the evaluation metrics, ensuring the solution is both accurate and scalable.

**CHAPTER 4: EXPERIMENTS AND RESULTS ANALYSIS**

**4.1 Model Training**

The training process involves splitting the data into training and testing sets. Hyperparameters for each model are optimized using grid search.

The model training phase is essential for enabling the prediction of ad clicks, involving splitting the dataset, optimizing hyperparameters, and training the chosen machine learning models. This process ensures that the models can learn patterns effectively and generalize well to unseen data.

**1. Dataset Splitting**

The dataset is divided into two subsets to facilitate training and evaluation:

* Training Set (80%): Used to train the model and identify patterns in the data.
* Testing Set (20%): Held back for evaluating the model's generalization performance.

A stratified sampling method is used to maintain the original class distribution across both sets, ensuring that the proportion of clicked and non-clicked ads remains consistent. This approach is particularly important in imbalanced datasets, which are common in ad click prediction.

**2. Hyperparameter Optimization**

Hyperparameters are fine-tuned to enhance model performance using a grid search approach combined with cross-validation (e.g., 5-fold cross-validation).

**1.Key Parameters:**

* Logistic Regression: Regularization strength (C) to control overfitting.
* Random Forest: Number of trees (n\_estimators), maximum tree depth (max\_depth), and maximum features (max\_features).
* Neural Networks: Number of layers, neurons per layer, learning rate, and dropout rate for regularization.

**2.Optimization Process:**

* Train models using different hyperparameter combinations.
* Evaluate performance metrics, such as accuracy, precision, recall, and F1-score, for each combination.
* Select the optimal set of hyperparameters that maximizes performance on the validation set.

**3. Training and Regularization**

The models are trained iteratively, ensuring robust learning while avoiding overfitting. Key strategies include:

* **Regularization:** Applying L2 regularization for Logistic Regression and dropout layers for Neural Networks.
* **Early Stopping:** For Neural Networks, halting training when validation loss stops improving.
* **Ensemble Learning:** Random Forest inherently reduces overfitting by aggregating predictions from multiple trees.

Data preprocessing steps, such as normalization and feature encoding, are incorporated into the pipeline to standardize inputs across all models.

**4.2 Evaluation**

Performance metrics such as accuracy, precision, and recall are used to assess model effectiveness.

The evaluation phase assesses the effectiveness of the trained models in predicting ad clicks. Using appropriate performance metrics ensures a comprehensive understanding of how well the models generalize to unseen data and meet the project's objectives. Given the potential imbalance in the dataset (fewer clicks compared to non-clicks), a combination of metrics is used to provide a balanced evaluation.

**1. Performance Metrics**

The following metrics are employed to evaluate model performance:

1. **Accuracy:**

Accuracy measures the proportion of predictions that were correct, including both clicks and non-clicks, out of the total predictions made. While accuracy provides a general performance overview, it may be misleading in imbalanced datasets since the majority class (non-clicks) can dominate predictions.

1. **Precision:**

Precision calculates the proportion of ads predicted as clicked that were actually clicked. A high precision score indicates a low rate of false positives, which is crucial for avoiding unnecessary ad targeting.

1. **Recall (Sensitivity):**

Recall measures the proportion of actual clicked ads that the model correctly identified. A high recall score ensures the model captures most of the actual clicks, minimizing missed opportunities for engagement.

1. **F1-Score:**

The F1-score is the harmonic mean of precision and recall, providing a balanced measure that accounts for both false positives and false negatives. This metric is especially useful in cases of imbalanced datasets where precision and recall may conflict.

1. **ROC-AUC (Receiver Operating Characteristic - Area Under the Curve):**

This metric evaluates the model's ability to distinguish between clicked and non-clicked ads across various prediction thresholds. A higher AUC score indicates better classification performance and suggests the model is more effective at identifying true clicks without being influenced by the threshold setting.

1. **Log Loss:**

Log loss assesses the accuracy of the predicted probabilities by penalizing incorrect predictions, especially when the model is highly confident about them. A lower log loss value indicates more reliable probabilistic predictions.

**2. Evaluation Process**

1. **Testing on Held-Out Data:**

The trained models are evaluated on the testing set, which was not seen during training, to measure their generalization capabilities.

1. **Comparison Across Models:**

Each model’s performance is compared using the selected metrics, enabling the identification of the most effective algorithm.

1. **Cross-Validation Results:**

Metrics from cross-validation during training are reviewed to ensure consistency between training and testing performance, indicating the absence of overfitting.

1. **Threshold Tuning:**

For probabilistic models like Logistic Regression and Neural Networks, the decision threshold for predicting a click is adjusted to optimize the trade-off between precision and recall based on the specific use case.

**3. Interpretation of Results**

The metrics provide actionable insights:

* High accuracy but low precision indicates that the model tends to over-predict clicks, resulting in false positives.
* High precision but low recall suggests that the model is cautious in predicting clicks, potentially missing real opportunities.
* A balanced F1-score reflects the model's ability to manage imbalanced data effectively, striking a compromise between precision and recall.

By analyzing these metrics, the best-performing model is selected, ensuring it aligns with business objectives like maximizing ad click-through rates while minimizing wasted ad spend.

**CHAPTER 5: CONCLUSION AND FUTURE WORK**

**5.1 Conclusion**

This study demonstrates that machine learning models can effectively predict ad clicks, with Random Forest and Neural Networks showing the highest performance.

**5.2 Future Work**

Further research could explore deep learning models and real-time click prediction for enhanced accuracy and scalability.

**REFERENCES**

*List of references used in the report.*

**APPENDIX**

*Any additional data or code snippets.*

**PUBLICATIONS**

*Any related publications or presentations.*