SUMMER INTERNSHIP REPORT

(May – July 2024)

Akash K V
420PH5021
Final-year
Integrated MSc
kvakash180@gmail.com

Title of the Project:

Web Scrapping and YouTube Ad View Prediction

Supervisor: Mr. Kashish Kumar

Platform where Internship Work was Done: **Jumia and Internshipstudio**

Area of Research:

Machine Learning and Data Analytics in Digital Marketing

The research area for this project lies at the intersection of machine learning, data analytics, and digital marketing. Machine learning, a subset of artificial intelligence, involves the development of algorithms that allow computers to learn patterns and make decisions based on data. Data analytics involves examining, cleaning, transforming, and modeling data to discover useful information, draw conclusions, and support decision-making.

Machine Learning in Predictive Analytics:
Predictive analytics uses historical data to predict future outcomes. In the context of digital marketing, it helps businesses forecast customer behavior, optimize marketing strategies, and improve customer engagement. By analyzing patterns and relationships in data, machine learning algorithms can predict key metrics, such as ad views, click-through rates, and conversion rates. These predictions are crucial for businesses that depend on online advertising revenue, as they allow for better budget allocation and targeting strategies.

Relevance to YouTube Ad View Prediction:
YouTube, one of the largest video-sharing platforms, serves as a critical channel for advertisers. Advertisers pay content creators based on the number of ad views and clicks their videos generate. Predicting ad views accurately helps advertisers optimize their ad placements, maximize reach, and improve return on investment (ROI). Machine learning models can analyze past video performance metrics and predict future ad views, enabling advertisers to target the right audience with tailored content.

Tools and Techniques in This Research:

1. **Supervised Learning:** This approach was primarily used, where historical data labeled with the number of ad views was used to train regression models. These models learn from the input features (like views, likes, comments) to predict the target variable (ad views).
2. **Data Visualization:** Techniques such as heatmaps, scatter plots, and histograms were used to explore the relationships and correlations among different features. Visualization helps in understanding data patterns, identifying outliers, and selecting important features.
3. **Feature Engineering:** The transformation of raw data into features that better represent the underlying problem to improve the predictive power of machine learning models. For example, converting published dates into numerical values, analyzing the duration of videos, and normalizing features.
4. **Model Evaluation Metrics:** Metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared score were used to evaluate model performance. These metrics provide insights into how well the model can predict ad views and its ability to generalize to new, unseen data.

Emerging Trends and Innovations:

- **Deep Learning:** With the increasing availability of computational power and large datasets, deep learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are being used for more complex prediction tasks. These models can capture non-linear patterns and dependencies better than traditional models.
- **Real-Time Analytics:** The demand for real-time data analysis is growing. Real-time prediction models can process live data streams and provide instant insights, which is valuable for time-sensitive advertising campaigns.
- **Integration with Big Data:** The use of big data technologies enables the processing and analysis of vast amounts of data from multiple sources. This integration enhances the predictive accuracy by considering a broader range of variables and historical data.
- **Ethical AI and Data Privacy:** As predictive analytics becomes more prevalent, concerns around data privacy and ethical AI practices are increasing. Ensuring that machine learning models are transparent, explainable, and comply with data privacy regulations is becoming a critical area of research.

Definition of the Problem:

The primary objective of this project was to predict the number of ad views on YouTube videos based on various video metrics, such as views, likes, dislikes, comments, published date, duration, and category. This prediction is crucial for YouTube advertisers, who pay content creators based on the number of ad views and clicks generated by their videos. Accurate ad view predictions enable advertisers to optimize their campaigns.

In addition to the YouTube ad view prediction, the project also involved web scraping data from the e-commerce website Jumia. Web scraping is a technique used to extract large amounts of data from websites automatically. This process provided hands-on experience in collecting, processing, and analyzing real-time data from an online retail platform. The data gathered from Jumia included product listings, prices, ratings, reviews, and other relevant details. The aim was to demonstrate how data from different sources, such as video platforms and e-commerce sites, can be used for analytics and predictive modeling.

Combining YouTube ad view prediction with web scraping of e-commerce data reflects the growing trend of using diverse data sources to gain insights into consumer behavior and market trends. By learning how to handle and analyze data from different domains, the project equipped me with valuable skills in data collection, preprocessing, and machine learning. This comprehensive approach ensures that predictive models are built with high accuracy and are capable of generalizing across various types of data.

Methodology (Computational):

The project followed a computational approach involving data analysis, machine learning model training, and evaluation. The methodology included the following steps:

1. **Data Import and Exploration:** The dataset was imported using Pandas, and initial exploratory analysis was conducted to understand the data shape, types, and distribution of each attribute.
2. **Data Visualization:** Visualizations such as heatmaps and distribution plots were created using Matplotlib and Seaborn to study relationships between different attributes and the target variable (ad views). This step provided insights into correlations and patterns in the data.
3. **Data Cleaning and Preprocessing:** Missing values were handled, and non-numeric attributes (e.g., published date, duration) were transformed into numerical values. Data normalization was performed to standardize the scale of input features.
4. **Data Splitting:** The dataset was split into training, validation, and test sets to ensure proper model evaluation and prevent overfitting. The standard ratio used was 70% for training, 15% for validation, and 15% for testing.

5. **Model Training:** Various regression models were trained on the dataset. Linear Regression and SVR models were implemented first as baselines, followed by more complex models like Decision Tree and Random Forest Regressors. Lastly, an ANN model was built and trained using Keras.
6. **Model Evaluation:** Each model's performance was evaluated based on error metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared score. Cross-validation was used to assess model generalization.
7. **Model Selection and Prediction:** The best-performing model, based on its accuracy and ability to generalize, was selected to make predictions on the test set. The model was saved for future use.

New Knowledge/Skills Learnt:

During this internship, I gained hands-on experience with several machine learning techniques and data analysis tools:

- Mastery of Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn for data manipulation, visualization, and analysis.
- Practical knowledge of machine learning models, including linear regression, SVR, decision trees, random forests, and neural networks.
- Skills in data cleaning, normalization, and transformation to prepare data for model training.
- Experience with Keras for building and training neural networks.
- Understanding of model evaluation metrics and techniques for preventing overfitting.
- Insights into the application of machine learning in digital marketing and advertising.

Related work Around the World:

Globally, machine learning and data analytics are extensively used in digital marketing to optimize advertising campaigns, predict user behavior, and personalize content. Companies like Google, Facebook, and Amazon leverage these techniques to analyze massive datasets and improve ad targeting, maximizing ad revenue and user engagement. Predictive analytics in advertising is a rapidly growing field, with continuous research focusing on developing more accurate and efficient models. Advances in neural networks and deep learning further enhance the ability to handle complex data structures and improve prediction capabilities.

Results Obtained:

The project involved training several machine learning models to predict the number of ad views. The models used include:

1. **Linear Regression:** This model served as a baseline for comparing other models. It provided initial predictions but showed limitations in handling non-linear relationships between input features and the target variable.
2. **Support Vector Regressor (SVR):** This model improved the prediction accuracy compared to linear regression by capturing complex patterns in the data. However, it required careful tuning of parameters like kernel type and regularization.
3. **Decision Tree Regressor:** This model offered higher accuracy by splitting the data into branches based on the input features. It managed to capture non-linear relationships effectively but was prone to overfitting on training data.
4. **Random Forest Regressor:** By using an ensemble of decision trees, this model provided better generalization and reduced overfitting. It achieved one of the best results in terms of prediction accuracy and error metrics.
5. **Artificial Neural Network (ANN):** Implemented using Keras, the ANN model was trained with multiple layers and various hyperparameters. It showed promising results by capturing intricate patterns in the data, and after experimentation with different architectures, it provided comparable accuracy to the Random Forest model.

Future Scope of Work:

The project offers several avenues for future research and improvement:

1. **Feature Engineering:** Exploring additional features, such as video content type, sentiment analysis of comments, and user engagement metrics, could improve prediction accuracy.
2. **Advanced Models:** Implementing more sophisticated deep learning models, such as Convolutional Neural Networks (CNNs) for video content analysis or Recurrent Neural Networks (RNNs) for time-series prediction based on published date trends.
3. **Real-Time Prediction:** Developing models capable of real-time ad view prediction to assist advertisers in making instantaneous decisions during ad campaigns.
4. **Integration with Web Scraping:** Enhancing the dataset by integrating real-time data collection through web scraping techniques to gather up-to-date metrics from YouTube and other social media platforms.
5. **Scalability:** Optimizing the models to handle larger datasets, potentially integrating data from multiple social media platforms to generalize the prediction models across different video content types and user bases.