**Industrial Internship Report on**

**YouTube Adview Prediction**

**Prepared by**

**JATIN**

|  |
| --- |
| *Executive Summary* |
| This report provides details of the Industrial Internship provided by upskill Campus and The IoT Academy in collaboration with Industrial Partner UniConverge Technologies Pvt Ltd (UCT).  This internship was focused on a project/problem statement provided by UCT. We had to finish the project including the report in 6 weeks’ time.  My project was **YouTube Adview Prediction.**  This internship gave me a very good opportunity to get exposure to Industrial problems and design/implement solution for that. It was an overall great experience to have this internship. |

**TABLE OF CONTENTS**

[1 Preface 3](#_Toc139702806)

[2 Introduction 4](#_Toc139702807)

[2.1 About UniConverge Technologies Pvt Ltd 4](#_Toc139702808)

[2.2 About upskill Campus 8](#_Toc139702809)

[2.3 The IOT Academy 9](#_Toc139702810)

[2.4 Objectives 9](#_Toc139702811)

[3 Problem Statement 1](#_Toc139702813)0

[4 Existing and Proposed solution 1](#_Toc139702814)1

5. Proposed Design/ Model ……………………………………………………………………………………………………………… 13

[6 Performance Test 1](#_Toc139702819)7

[6.1 Test Plan/ Test Cases 17](#_Toc139702820)

[6.2 Test Procedure 1](#_Toc139702821)8

[6.3 Performance Outcome 1](#_Toc139702822)9

[7 My learnings](#_Toc139702823) 20

[8 Future work scope](#_Toc139702824) 21

# Preface

This six-week internship was a significant milestone in my career development. It provided an opportunity to work on a practical problem that required the application of advanced machine learning techniques. The primary goal was to develop a predictive model that could estimate YouTube adview counts based on various video features. This report documents the entire process, from problem identification to the final solution.

The project was meticulously planned to cover all aspects of machine learning, including data preprocessing, model selection, and performance evaluation. I had the opportunity to explore different machine learning algorithms and understand their strengths and limitations in a real-world scenario. This internship also highlighted the importance of industrial internships in bridging the gap between academic knowledge and industry requirements.



I would like to thank UniConverge Technologies Pvt Ltd (UCT) for providing this opportunity and upskill Campus (USC) for their continuous support throughout the internship. I also extend my gratitude to my mentors, colleagues, and peers who contributed to this successful learning experience. To my juniors and peers, I encourage you to take up similar opportunities that allow you to apply your knowledge in real-world settings.

# Introduction

## About UniConverge Technologies Pvt Ltd

A company established in 2013 and working in Digital Transformation domain and providing Industrial solutions with prime focus on sustainability and RoI.

For developing its products and solutions it is leveraging various**Cutting Edge Technologies e.g. Internet of Things (IoT), Cyber Security, Cloud computing (AWS, Azure), Machine Learning, Communication Technologies (4G/5G/LoRaWAN), Java Full Stack, Python, Front end**etc.



1. UCT IoT Platform **(****)**

**UCT Insight** is an IOT platform designed for quick deployment of IOT applications on the same time providing valuable “insight” for your process/business. It has been built in Java for backend and ReactJS for Front end. It has support for MySQL and various NoSql Databases.

* It enables device connectivity via industry standard IoT protocols - MQTT, CoAP, HTTP, Modbus TCP, OPC UA
* It supports both cloud and on-premises deployments.

It has features to  
• Build Your own dashboard  
• Analytics and Reporting  
• Alert and Notification  
• Integration with third party application(Power BI, SAP, ERP)  
• Rule Engine

1. **Smart Factory Platform (****)**

Factory watch is a platform for smart factory needs.

It provides Users/ Factory

* with a scalable solution for their Production and asset monitoring
* OEE and predictive maintenance solution scaling up to digital twin for your assets.
* to unleased the true potential of the data that their machines are generating and helps to identify the KPIs and also improve them.
* A modular architecture that allows users to choose the service that they what to start and then can scale to more complex solutions as per their demands.

Its unique SaaS model helps users to save time, cost and money. 



1.  based Solution

UCT is one of the early adopters of LoRAWAN teschnology and providing solution in Agritech, Smart cities, Industrial Monitoring, Smart Street Light, Smart Water/ Gas/ Electricity metering solutions etc.

1. Predictive Maintenance

UCT is providing Industrial Machine health monitoring and Predictive maintenance solution leveraging Embedded system, Industrial IoT and Machine Learning Technologies by finding Remaining useful life time of various Machines used in production process.



## About upskill Campus (USC)

upskill Campus along with The IoT Academy and in association with Uniconverge technologies has facilitated the smooth execution of the complete internship process.

USC is a career development platform that delivers **personalized executive coaching** in a more affordable, scalable and measurable way.



Seeing need of upskilling in self paced manner along-with additional support services e.g. Internship, projects, interaction with Industry experts, Career growth Services

<https://www.upskillcampus.com/>

upSkill Campus aiming to upskill 1 million learners in next 5 year





## The IoT Academy

The IoT academy is EdTech Division of UCT that is running long executive certification programs in collaboration with EICT Academy, IITK, IITR and IITG in multiple domains.

## Objectives of this Internship program

The objective for this internship program was to

 ☛ get practical experience of working in the industry.

 ☛ to solve real world problems.

 ☛ to have improved job prospects.

 ☛ to have Improved understanding of our field and its applications.

 ☛ to have Personal growth like better communication and problem solving.

# 3. Problem Statement

In the assigned problem statement

The goal of this project was to develop a predictive model for estimating the number of ad views a YouTube video might receive. The problem is complex due to the wide range of factors influencing adview counts, including video metrics like views, likes, dislikes, comments, and the video category. The challenge was to preprocess the raw data, extract meaningful features, and build a robust machine learning model that could accurately predict adview counts.

The significance of this problem lies in its practical application in the digital marketing industry, where accurate predictions of adview counts can inform advertising strategies, budget allocation, and content creation. The solution required careful consideration of various data preprocessing techniques, model selection, and evaluation metrics to ensure the predictions were reliable and actionable.

**4. Existing and Proposed solution**

**Existing Solutions**

Existing solutions for predicting YouTube adview counts often rely on simple statistical models or heuristic approaches. These methods typically use linear regression or basic correlation analysis to estimate adview counts based on a few selected features. However, these models often fail to capture the complex, non-linear relationships between video features and ad views, leading to suboptimal predictions.

**The limitations of existing solutions include:**

**Oversimplification**: Basic models may not account for interactions between features, leading to inaccurate predictions.

**Limited Feature Use**: Many existing models do not fully utilize all available features, resulting in loss of valuable information.

**Poor Generalization**: Models trained on a limited dataset may not generalize well to new data, reducing their practical applicability.

**Proposed Solution**

The proposed solution involves using advanced machine learning algorithms to build a more accurate and robust predictive model.

**The approach includes:**

**Data Preprocessing**: Comprehensive preprocessing of the dataset to handle missing values, encode categorical data, and normalize numerical features.

**Feature Engineering**: Creating new features that better represent the underlying patterns in the data, such as interaction terms or transformed variables.

**Model Selection**: Training and comparing multiple machine learning models, including linear regression, decision trees, random forests, support vector machines, and artificial neural networks.

**Model Evaluation**: Using cross-validation and performance metrics such as mean absolute error (MAE) and root mean squared error (RMSE) to evaluate the models’ accuracy and robustness.

**Value Addition:** The proposed model aims to provide more accurate predictions, which can be used by digital marketers to optimize their ad placement strategies and maximize ROI.

## Code submission (Github link): [Jatin0804/upskillcampus](https://github.com/Jatin0804/upskillcampus)

## Report submission (Github link) : [upskillcampus/Report.docx](https://github.com/Jatin0804/upskillcampus/blob/main/Report.docx)

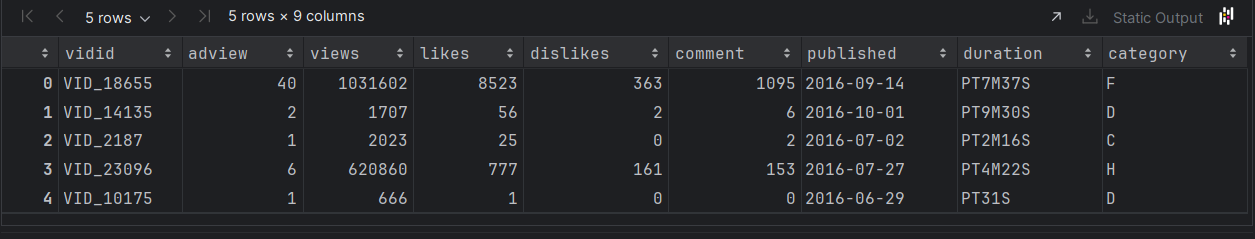
## 

**5. Proposed Design/Model**

**Data Preprocessing**

Data preprocessing is a critical step in building a successful machine learning model. In this project, the preprocessing phase included:

* **Data Cleaning:** Removing any duplicate or irrelevant records and addressing missing values through imputation or deletion.
* **Categorical Encoding:** Converting categorical variables, such as video category, into numerical format using techniques like one-hot encoding or label encoding.
* **Feature Scaling:** Normalizing numerical features to ensure that all variables contribute equally to the model's predictions. This was done using standardization or min-max scaling.



**Feature Engineering**

Feature engineering involved creating new features from the existing data that could improve the model’s performance. For example:

* **Interaction Features:** Creating interaction terms between views and likes to capture their combined effect on adview counts.
* **Log Transformation:** Applying a logarithmic transformation to highly skewed variables, such as the number of views, to reduce the impact of outliers.

**Model Training**

Multiple machine learning models were trained on the processed dataset. The models included:

* **Linear Regression:** A basic model that serves as a benchmark, predicting adviews based on a linear combination of input features.
* **Decision Trees:** A non-linear model that splits the data based on feature values to make predictions.
* **Random Forests:** An ensemble model that improves the robustness and accuracy of decision trees by averaging multiple trees.
* **Support Vector Machines (SVM):** A model that finds the hyperplane that best separates the data into different classes, used here for regression.
* **Artificial Neural Networks (ANN):** A deep learning model that uses multiple layers to capture complex patterns in the data.

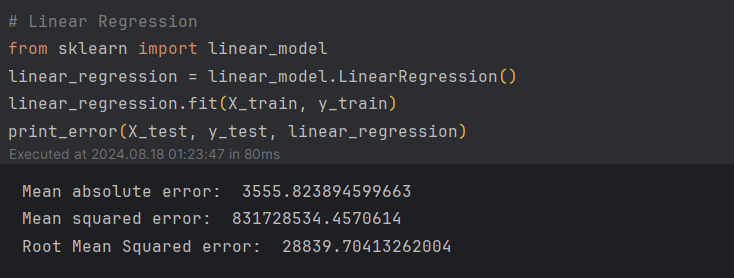
**Model Evaluation**

The models were evaluated using cross-validation and performance metrics:

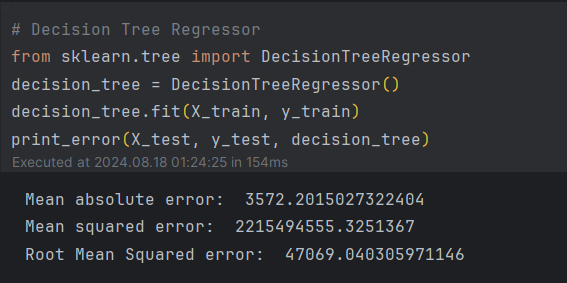
* **Mean Absolute Error (MAE):** Measures the average absolute difference between the predicted and actual adview counts.
* **Root Mean Squared Error (RMSE):** Penalizes larger errors more heavily than MAE, providing a more sensitive measure of model performance.
* **R-Squared (R²):** Indicates the proportion of variance in the dependent variable that is predictable from the independent variables.

The best-performing model was selected based on these metrics and was further fine-tuned to improve its accuracy.

**Linear Regression**



**Decision Tree Regressor**



**Random Forest Regressor**

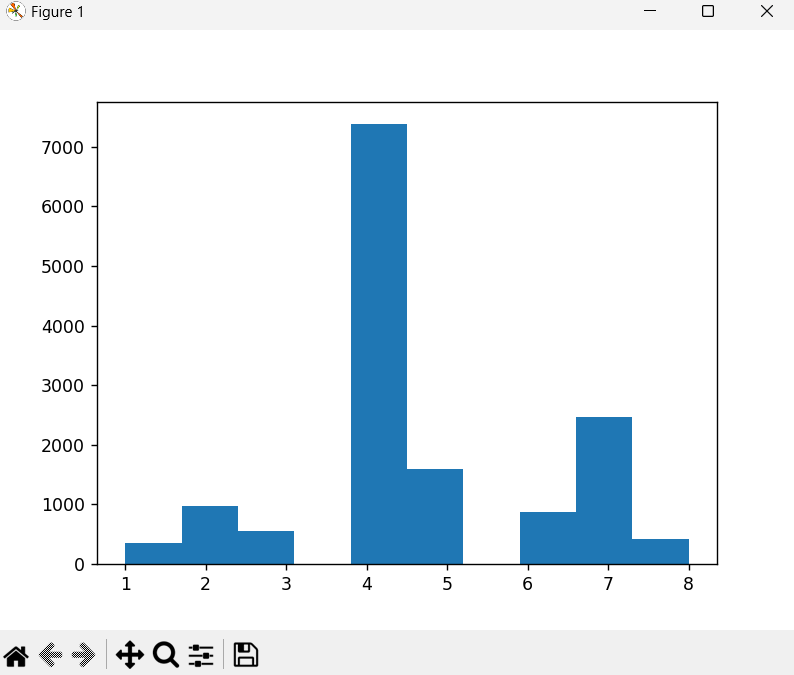


# Performance Test

1. **Test Plan/Test Cases**

The test plan was designed to evaluate the performance of the models on the test dataset. The key test cases included:

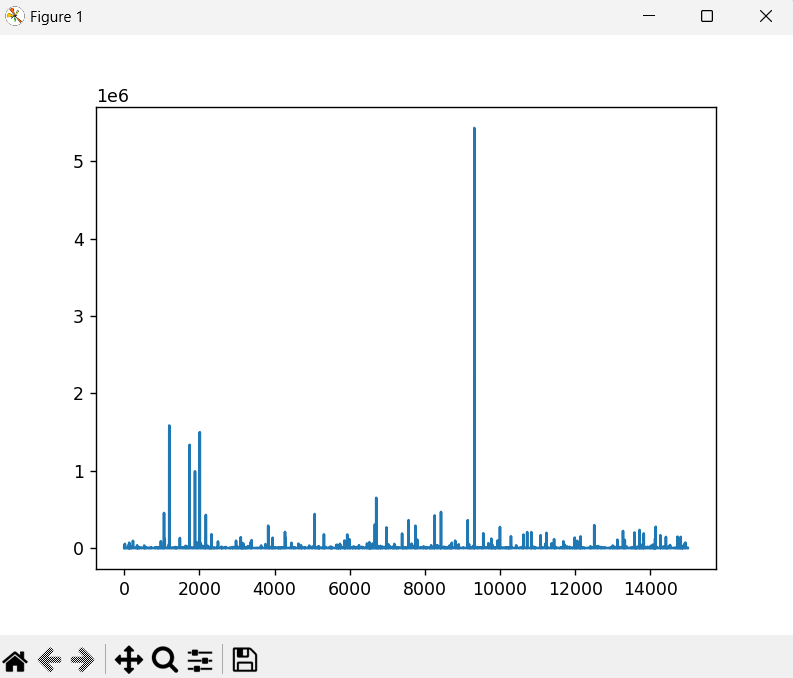
* **Model Accuracy:** Ensuring that the predictions are within an acceptable range of the actual adview counts.
* **Generalization:** Testing the model on unseen data to verify that it generalizes well beyond the training dataset.
* **Robustness:** Evaluating the model's performance under different conditions, such as varying the feature set or introducing noise into the data.

****

1. **Test Procedure**

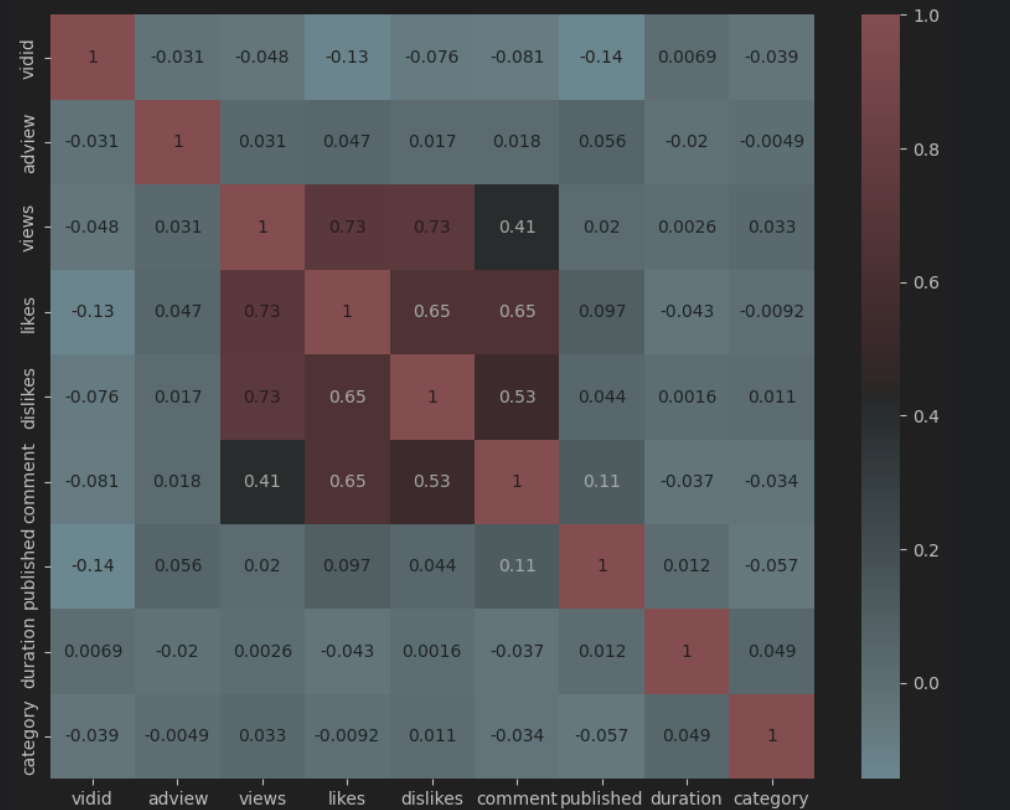
The test procedure followed a systematic approach:

* 1. **Data Split:** The dataset was divided into training and test sets, typically in a 70:30 ratio.
  2. **Model Training:** Each model was trained on the training set, using cross-validation to optimize hyperparameters and prevent overfitting.
  3. **Performance Evaluation:** The models were evaluated on the test set using the MAE, RMSE, and R² metrics. The results were compared to identify the best-performing model.
  4. **Model Tuning:** The selected model was further fine-tuned by adjusting hyperparameters and retraining to achieve the best possible performance.



1. **Performance Outcome**

The random forest regressor emerged as the best-performing model with an RMSE of approximately 0.23 million adviews. This model demonstrated a good balance between accuracy and generalization, making it suitable for predicting adview counts in a real-world setting. The neural network also performed well, particularly in capturing complex non-linear relationships in the data, but required more computational resources and fine-tuning.



# My learnings

This internship was a transformative experience that enhanced both my technical skills and my understanding of machine learning applications. Key learnings include:

* **Data Preprocessing:** I gained a deeper understanding of the importance of data cleaning, feature engineering, and the impact of different preprocessing techniques on model performance.
* **Model Selection:** The experience taught me how to choose the right machine learning model based on the problem at hand, considering factors like model complexity, interpretability, and computational efficiency.
* **Performance Evaluation:** I learned how to evaluate models using appropriate metrics and cross-validation techniques to ensure that the model is not just accurate but also generalizes well to new data.
* **Problem-Solving:** The project improved my problem-solving skills, particularly in handling challenges like missing data, feature selection, and model tuning.
* **Collaboration:** Working in a collaborative environment with mentors and peers helped me improve my communication and teamwork skills, which are crucial for success in any professional setting.

# Future work scope

While the project achieved its primary objective, there are several areas for future improvement:

* **Hyperparameter Optimization:** Further fine-tuning of the model's hyperparameters could improve its predictive accuracy.
* **Feature Expansion:** Incorporating additional features, such as video tags, channel statistics, or external factors like trends and seasonality, could enhance the model's performance.
* **Advanced Techniques:** Exploring advanced machine learning techniques, such as ensemble learning with stacking or boosting, could provide even better results.
* **Deployment:** The final model could be deployed as a web application or integrated into a digital marketing platform to provide real-time adview predictions.
* **Real-Time Data:** Incorporating real-time data into the model could make the predictions more dynamic and responsive to current trends.
* **User Interface:** Developing a user-friendly interface for the model would make it accessible to non-technical users, allowing marketers to easily input data and obtain predictions.