INDUSTRIAL TRAINING REPORT ON PROJECT “YOUTUBE ADVIEW PREDICTION” MACHINE LEARNING AT INTERNSHIP STUDIO

Name: Trisha Sahu

Class: BE CSE 5th Semester

Roll No: SG19357

ORGANISATION PROFILE

Internship Studio is a platform developed to help students build their profiles by providing them the right exposure to develop the required skills in the respective domain.

❖ Encouraging students to work on projects & learn from the professionals.
❖ Infusing a learning spirit through the best of best mentorship.
❖ Filling the gap between bookish knowledge and practical knowledge by providing training + internship

Instructor: Kashish Kumar

TOPICS

Training Week 1
Introduction to Statistics
Summary Statistics
Probability
Permutations and Combinations
Discrete Probability Distributions
Continuous Probability Distributions
Inferential Statistics

Training Week 2
Basics of Python Programming
Advanced Python Programming
Python Libraries: Numpy
Python Libraries: Pandas
Python Libraries: Matplotlib

Training Week 3
Introduction to Machine Learning
Python Libraries: Sklearn
Linear Regression
Logistic Regression
Decision Tree and Random Forest
Ensemble Techniques
Naïve Bayes and SVM
Unsupervised Learning
Key ML Algorithms – KNN
Neural Network and Deep Learning

Internship Week 1
ML Internship Project Problem Statement

Internship Week 2
ML Internship Project Submission

Technology and Concepts

Machine Learning

In classic terms, machine learning is a type of artificial intelligence that enables self-learning from data and then applies that learning without the need for human intervention.

Linear Regression

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog).

There are two main types:

1. Simple Regression
2. Multiple Regression

Support Vector Machine

“Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate.

Decision Tree

Decision tree analysis involves making a tree-shaped diagram to chart out a course of action or a statistical probability analysis. It is used to break down complex problems or branches. Each branch of the decision tree could be a possible outcome.

Artificial Neural Network (ANN)

An artificial neural network (ANN) is the piece of a computing system designed to simulate the way the human brain analyzes and processes information. It is the foundation of artificial intelligence (AI) and solves problems that would prove impossible or difficult by human or statistical standards. ANNs have self-learning capabilities that enable them to produce better results as more data becomes available.

PROJECT PROBLEM STATEMENT

Youtube advertisers pay content creators based on adviews and clicks for the goods and services being marketed. They want to estimate the adview based on other metrics like comments, likes etc.

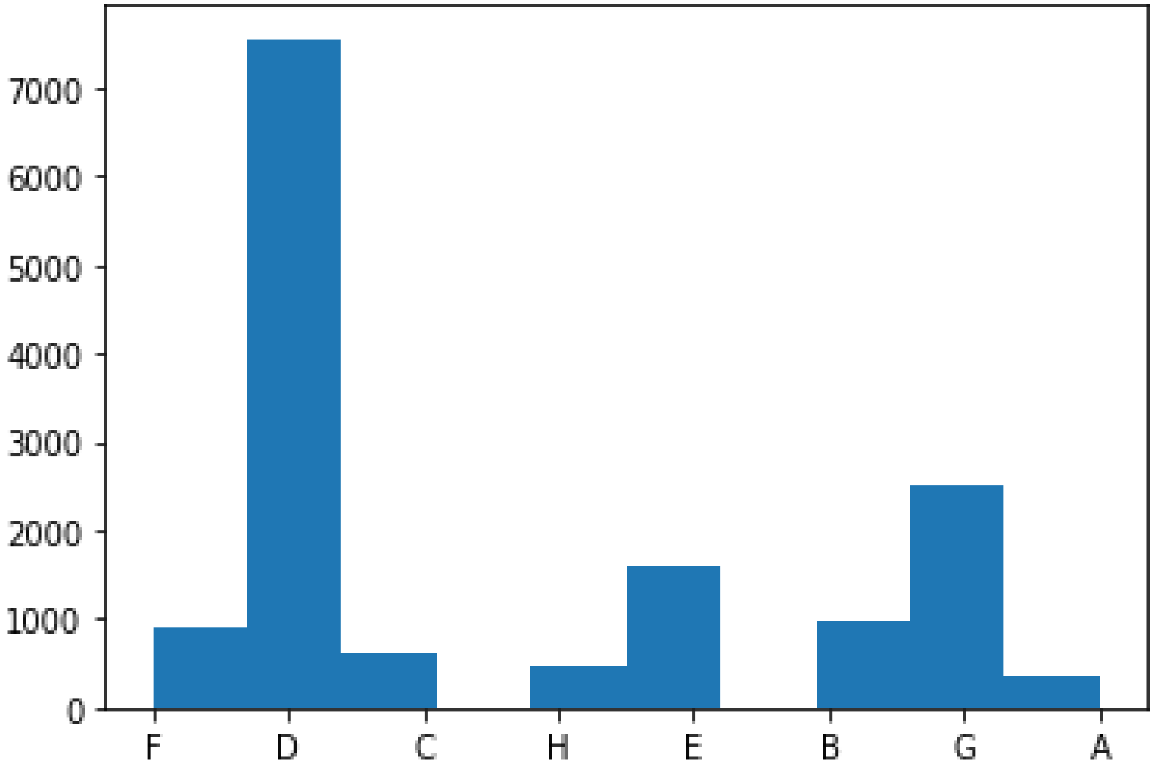
The problem statement is therefore to train various regression models and choose the best one to predict the number of adviews. The data needs to be refined and cleaned before feeding in the algorithms for better results.

Steps For Adview Prediction

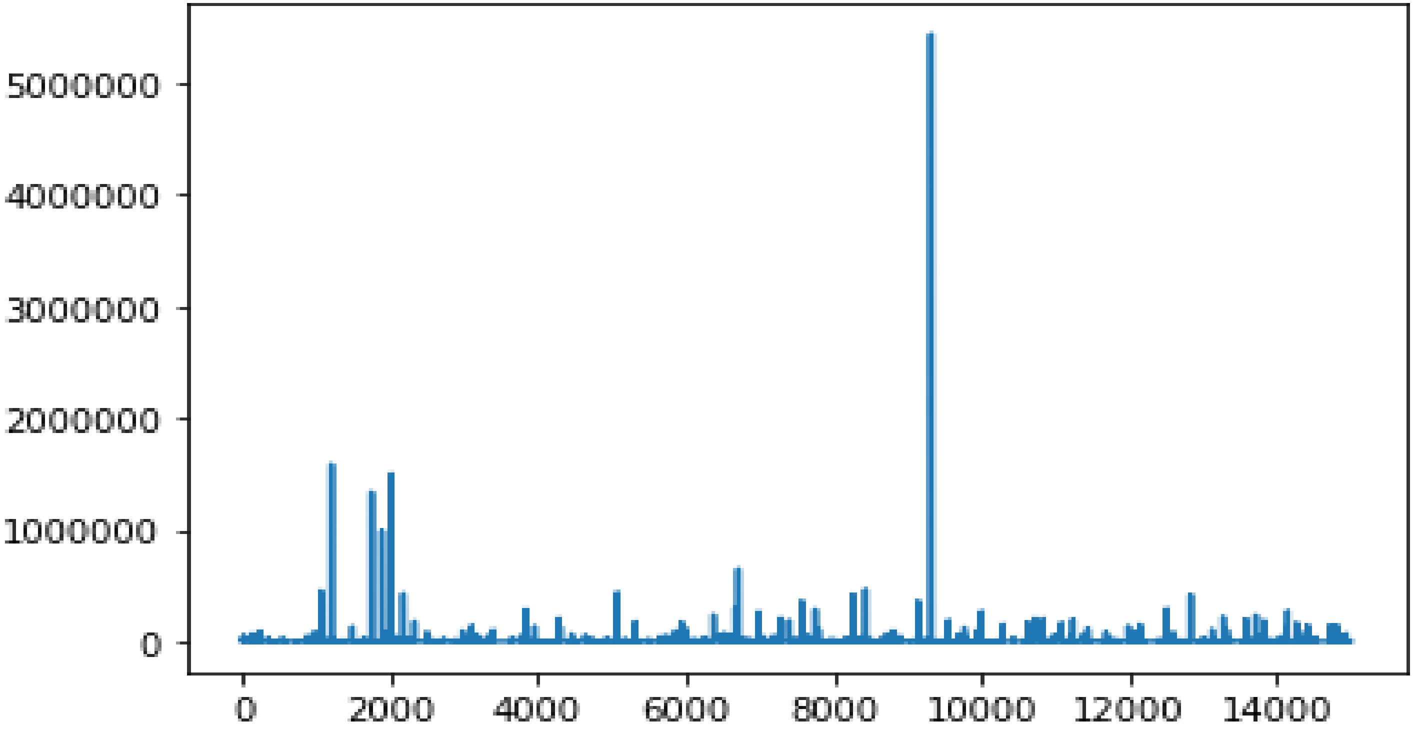
1. Import the datasets and libraries, check shape and datatype.
2. Visualise the dataset using plotting using heatmaps and plots. You can study data distributions for each attribute as well.
3. Clean the dataset by removing missing values and other things.
4. Transform attributes into numerical values and other necessary transformations
5. Normalise your data and split the data into training, validation and test set in the appropriate ratio.
6. Use linear regression, Support Vector Regressor for training and get errors.
7. Use Decision Tree Regressor and Random Forest Regressors.
8. Build an artificial neural network and train it with different layers and hyperparameters.

Visualization

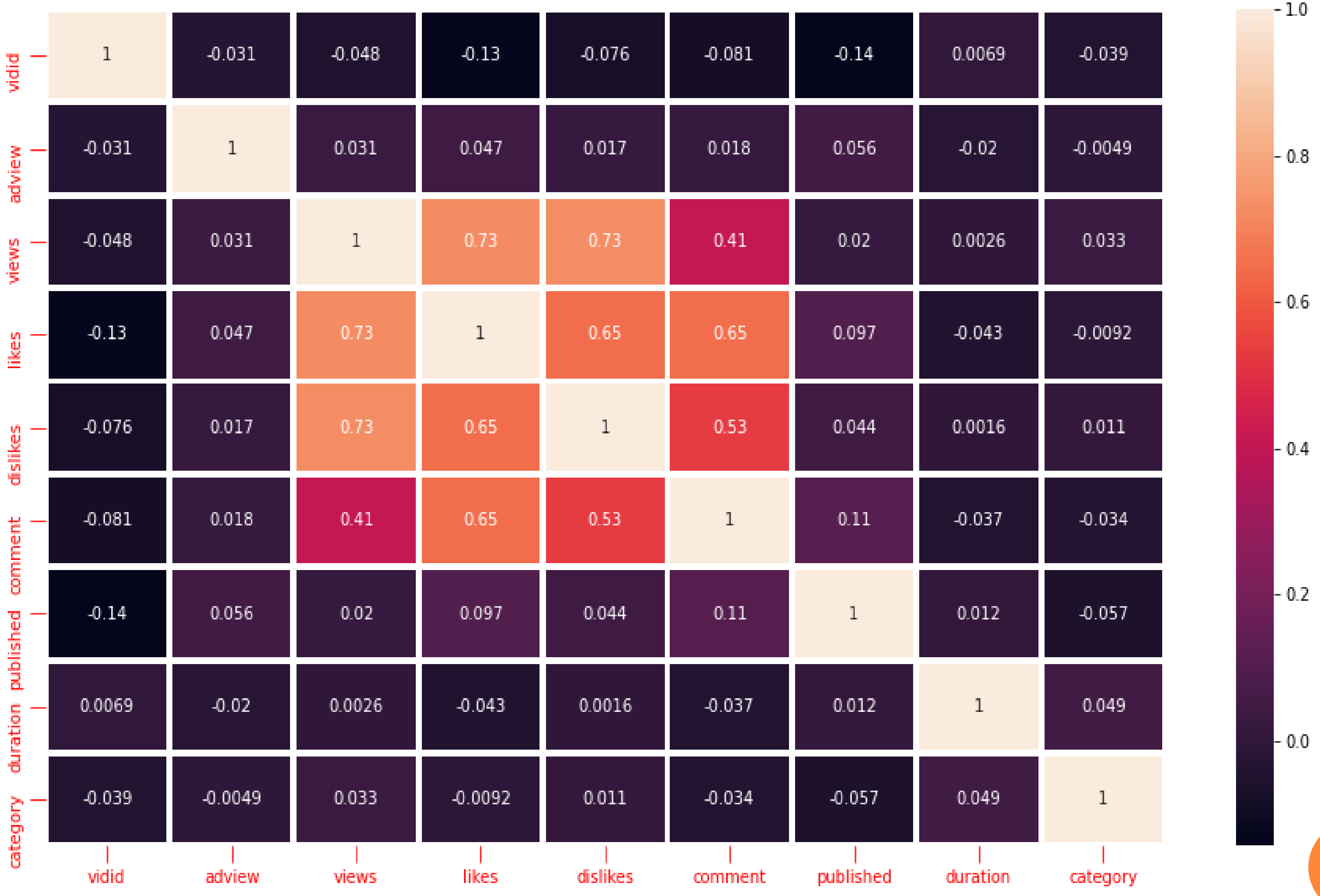
Histogram of “Category” column



Histogram of “adview” column



Heatmap which shows the co-relation of all columns with each other



	vidid	adview	views	likes	dislikes	comment	published	duration	category
vidid	1	-0.031	-0.048	-0.13	-0.076	-0.081	-0.14	0.0069	-0.039
adview	-0.031	1	0.031	0.047	0.017	0.018	0.056	-0.02	-0.0049
views	-0.048	0.031	1	0.73	0.73	0.41	0.02	0.0026	0.033
likes	-0.13	0.047	0.73	1	0.65	0.65	0.097	-0.043	-0.0092
dislikes	-0.076	0.017	0.73	0.65	1	0.53	0.044	0.0016	0.011
comment	-0.081	0.018	0.41	0.65	0.53	1	0.11	-0.037	-0.034
published	-0.14	0.056	0.02	0.097	0.044	0.11	1	0.012	-0.057
duration	0.0069	-0.02	0.0026	-0.043	0.0016	-0.037	0.012	1	0.049
category	-0.039	-0.0049	0.033	-0.0092	0.011	-0.034	-0.057	0.049	1

RESULT

Algorithm	Linear Regression	Random forest	Decision tree	Support vector machine	ANN
Mean Absolute Error	3707.378005824529	3274.6902966905504	3059.310792349727	3707.378005824529	3304.264894606637
Mean Squared Error	835663131.1210335	644433788.0361483	1226286165.4118853	835663131.1210335	829552666.7955565
Root Mean Squared Error	28907.83857573986	**25385.70046376795**	35018.3689713254	28907.83857573986	28801.95595433679

CONCLUSION

Best Model

From the training dataset by applying all algorithms for train the model, we found that **"Random Forest Regressor"** algorithm has less root mean squared error as compared to other algorithms. As we know model having less root mean squared error is more perfect.
So here for prediction of test dataset we use **"Random Forest" algorithm**.

THANK YOU !!!