ESTIMATION OF YOUTUBE EARNINGS BASED ON CHANNEL DEMOGRAPHICS

Project ID: P-32
Atharva Pansare, Ritwik Jog, Viraj Sanap
Department of Computer Science, North Carolina State University
Raleigh, NC 27695
[aspansar, rsjog, vasanap]@ncsu.edu

1 Background/Introduction

In recent years, YouTube has become a prominent platform for content creators, offering them an opportunity to monetize their content. This project focuses on the estimation of earnings generated by YouTube channels. Given the diversity of content and creators on the platform, understanding the factors that influence earnings is of academic interest. Our project employs data analysis and modeling to explore the variables and the underlying trends YouTube channel earnings, providing valuable insights for both creators and researchers in the field.

1.1 Project Idea

This project aims to predict YouTube channel earnings using data like subscribers and video views. We'll use the dataset available on Kaggle, preprocess it, and develop a predictive model to estimate earnings. By evaluating the model's performance and discussing key findings, we aim to provide insights into the factors influencing channel earnings.

1.2 Relevant Papers

[1] R. Shah et. al., "Cluster Based Analysis for Google YouTube Videos Viewers," 2020 International Conference on Computer Science, Engineering and Applications

[2] E. Ramalakshmi et. al., "YouTube Data Analysis and Prediction of Views and Categories." Ijraset Journal For Research in Applied Science and Engineering Technology

[3] N. Alias, G. elHadad et. al., "A Content Analysis in the Studies of YouTube in Selected Journals". November 2013 Procedia - Social and Behavioral Sciences

2 Method

2.1 Machine Learning Techniques Used

For our project, we have trained and compared the performances of the following different Machine Learning Regression Models on our dataset.

1. **Linear Regression:** Linear regression is a fundamental statistical method for modeling the relationship between a dependent variable and one or more independent variables. It aims to find a linear equation that best fits the data, enabling predictions and understanding of this relationship. The model assumes a straight-line relationship between variables, where the goal is to minimize the sum of the squared differences between observed data points and their corresponding predictions. Linear regression is widely used in fields like

economics, finance, and science for tasks such as forecasting, trend analysis, and determining the strength and direction of correlations. It's a simple yet powerful tool for making data-driven decisions.

2. **Ridge/Lasso Regression:** Ridge and Lasso regression are advanced linear regression techniques used in data analysis and machine learning to combat overfitting. Ridge adds a penalty term to the traditional least squares method, constraining the coefficients and reducing their magnitudes, which often results in improved model generalization. Lasso, on the other hand, combines the least squares method with a penalty term that can drive some coefficients to absolute zero, effectively selecting a subset of the most influential features. Both methods are valuable for managing multicollinearity and feature selection, with Ridge offering a smoother shrinkage and Lasso yielding sparser models.
3. **Support Vector Machine (SVM) Regression:** Support Vector Machine (SVM) Regression is a powerful machine learning technique used for predictive modeling. Unlike classification, SVM regression focuses on estimating continuous values rather than discrete classes. It works by finding a hyperplane that best fits the data, aiming to minimize the margin of error or the difference between the predicted and actual values. This model is particularly effective in cases where the relationship between input features and the target variable is nonlinear. SVM regression employs support vectors, which are data points closest to the regression line, to optimize its accuracy. It's a versatile tool for tasks like forecasting, and anomaly detection.
4. **Random Forest Regression:** Random Forest Regression is a powerful machine learning technique used for predictive modeling. It operates by constructing an ensemble of decision trees, each trained on a subset of the data. Instead of producing a single prediction, it generates multiple predictions and averages them for a more accurate and robust result. This method is particularly effective for handling complex, nonlinear relationships between variables and can be applied to various domains, such as finance, ecology, and healthcare, to make precise continuous predictions, making it a versatile tool in data-driven decision-making and analysis.
5. **Artificial Neural Netrworks (ANNs):** ANNs consist of interconnected nodes or neurons that process information. Each neuron performs a weighted sum of its inputs and applies an activation function to produce an output. ANNs are used for tasks such as pattern recognition, classification, regression, and more. They learn from data by adjusting their connection weights through training, enabling them to make predictions or decisions based on complex patterns. In this project, we have made use of a simple 5 layered Fully-Connected ANN with a total of 5,409 trainable parameters.

3 Plan and Experiment

3.1 Dataset Description

We have made use of the "Global YouTube Statistics 2023" dataset from Kaggle.com. The dataset contains information related to the top performing YouTube channels/creators such as their subscriber counts, channel genre, earnings, geographical location, etc. There were a total of 28 columns and close to 1000 records present in the dataset. The 28 features are as follows:

- rank: Position of the YouTube channel based on the number of subscribers
- Youtuber: Name of the YouTube channel
- subscribers: Number of subscribers to the channel
- video views: Total views across all videos on the channel
- category: Category or niche of the channel
- Title: Title of the YouTube channel
- uploads: Total number of videos uploaded on the channel
- Country: Country where the YouTube channel originates
- Abbreviation: Abbreviation of the country

- channel type: Type of the YouTube channel (e.g., individual, brand)
- video views rank: Ranking of the channel based on total video views
- country rank: Ranking of the channel based on the number of subscribers within its country
- channel type rank: Ranking of the channel based on its type (individual or brand)
- video views for the last 30 days: Total video views in the last 30 days
- lowest monthly earnings: Lowest estimated monthly earnings from the channel
- highest monthly earnings: Highest estimated monthly earnings from the channel
- lowest yearly earnings: Lowest estimated yearly earnings from the channel
- highest yearly earnings: Highest estimated yearly earnings from the channel
- subscribers for last 30 days: Number of new subscribers gained in the last 30 days
- created year: Year when the YouTube channel was created
- created month: Month when the YouTube channel was created
- created date: Exact date of the YouTube channel's creation
- Gross tertiary education enrollment (Percentage): Percentage of the population enrolled in tertiary education in the country
- Population: Total population of the country
- Unemployment rate: Unemployment rate in the country
- Urban population: Percentage of the population living in urban areas
- Latitude: Latitude coordinate of the country's location
- Longitude: Longitude coordinate of the country's location

3.2 Hypothesis

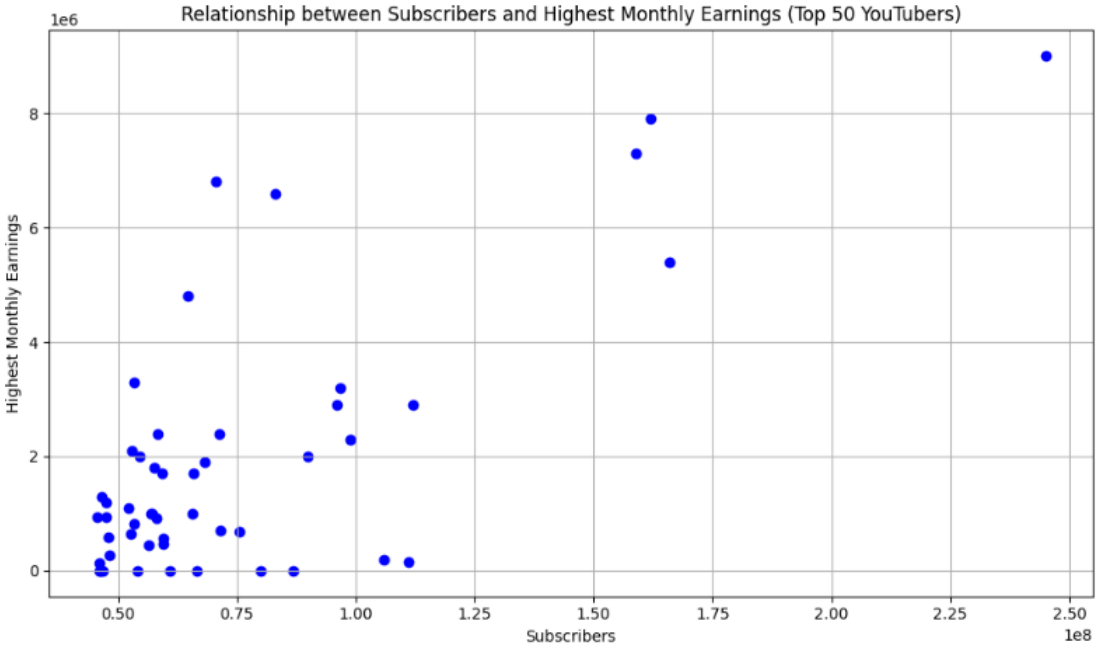
We believe that the annual earnings of a YouTube channel can be predicted and influenced by various parameters, including the number of subscribers, total video views, the country of the channel, and socio-economic factors related to that country.

- **Subscribers and Earnings:** We hypothesize that there exists a positive correlation between the number of subscribers a YouTube channel has and its annual earnings. Higher subscriber counts may attract more advertisers and sponsorship opportunities, contributing to increased revenue.
- **Video Views and Earnings:** We anticipate a positive relationship between the total number of video views on a channel and its annual earnings. Channels with higher viewership may have a broader audience, potentially attracting more advertising revenue.
- **Country of the Channel and Earnings:** Our hypothesis suggests that the country in which a YouTube channel operates may significantly impact its annual earnings. Factors such as the economic stability, advertising market, and consumer purchasing power in that country are expected to influence the revenue potential of the channel.
- **Socio-Economic Factors of the Country and Earnings:** We predict that socio-economic factors of the country, such as GDP per capita, internet penetration rate, and digital marketing infrastructure, may contribute to variations in the annual earnings of YouTube channels. Higher socio-economic indicators could lead to increased advertising spending and revenue potential.
- **Content Genre and Earnings:** While not explicitly mentioned, we may also hypothesize that the type of content produced by the channel could impact earnings. Certain genres may attract more lucrative sponsorship deals or targeted advertising.

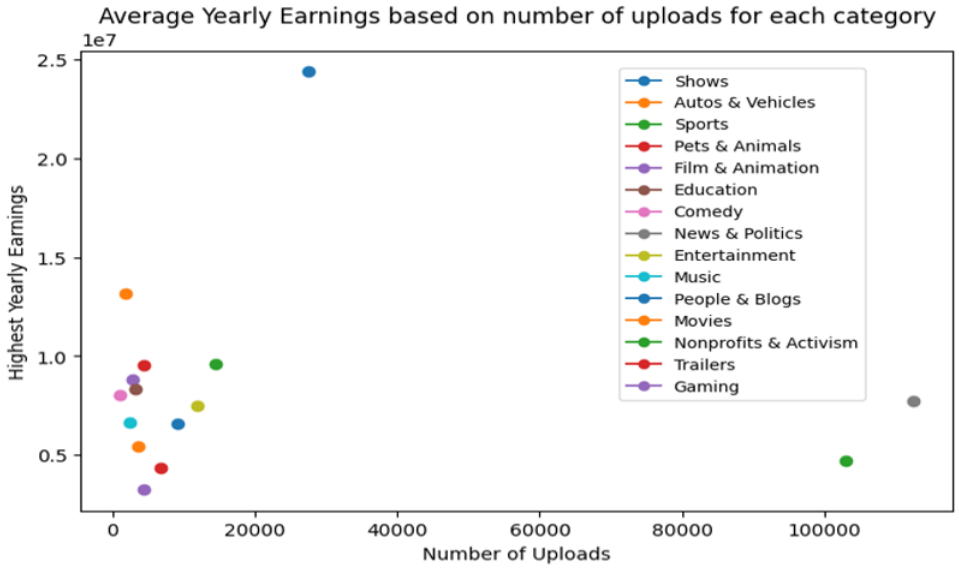
By conducting a comprehensive analysis of these parameters, we aim to uncover patterns and relationships that can provide insights into the factors influencing the annual earnings of YouTube channels.

3.3 Visualization and Initial Data Insights

Before carrying out any kind of data preprocessing or data cleaning operations, it is necessary to visualize our dataset in the form of plots and graphs to gain some initial insights and familiarity with it. Additionally, such graphs also help to uncover some hidden patterns or correlations between different variables.



In the above graph, it can be seen that there definitely is a positive correlation between the number of subscribers and the monthly earnings. However, that relationship doesn't seem linear in nature. Moreover, there are many cases where channels with less subscribers have higher earnings than channels with more subscribers. This indicates that the number of subscribers is not the only feature contributing towards the amount of earnings.



From the above figure, it can be inferred that, while the number of video uploads definitely affects the average earnings, the genre of the channel also greatly influences the outcome. "Non Profits

and Activism" related channels have a high number of video uploads but it can be seen that that doesn't result in high channel earnings. From this, we can conclude that the channel genre is also an important feature that contributes towards determining the earnings of a YouTube Channel per video upload.

3.4 Data Preprocessing

Data preprocessing plays a vital role in the Machine Learning pipeline. It helps us to identify and retain the most informative features. By processing the data we not only enhance the accuracy of the predictive model but it also lays a foundation for comprehensive exploratory data analysis. To fix this, we performed some basic data preprocessing which included the following steps:

- **Handling Missing Values**: We found that there were around 122 missing values for the columns related to the "Country" feature and all the features associated with that such as the "Population", "Unemployment Rate", "Literacy Rate", etc. Due to this, all of those 122 records were sparse and filling those missing values with some arbitrary values or with the mean of other records will worsen the quality of the dataset. So, we decided to drop such sparse records.
- **Text Vectorization**: The data contained some text which had special characters. This made processing the data tricky, and did not add any value to the dataset itself, so we used text vectorization to remove any special characters that were present in the data.
- **One-Hot Encoding**: Categorical variables were one-hot encoded to help with data analysis.
- **Feature Scaling**: The scales of different attributes differed greatly. For example, the number of uploads was usually a few hundred, but values such as earnings and video views were in the millions and billions respectively. Hence, we normalized our data to fit everything within a range of [0,1].
- **Outlier Detection and Removal**: Some records clearly stood out as Outliers. For example, channels such as YouTube Music did not make any money despite having some of the highest number of views and subscribers. We removed these outliers so that they did not interfere with our analysis.
- **Feature Engineering**: We found that our data contained some redundant columns, for example, country and country code. Hence, we dropped one of the columns since we did not stand to learn anything new from it. Similarly, we also created a few new columns such as Channel Age using existing columns (Year Created).
- **Dimensionality Reduction**: The initial dataset had 28 columns. To make our training more effective, we decided to reduce the dimensionality of our dataset using techniques such as Principal Component Analysis (PCA).

4 Results

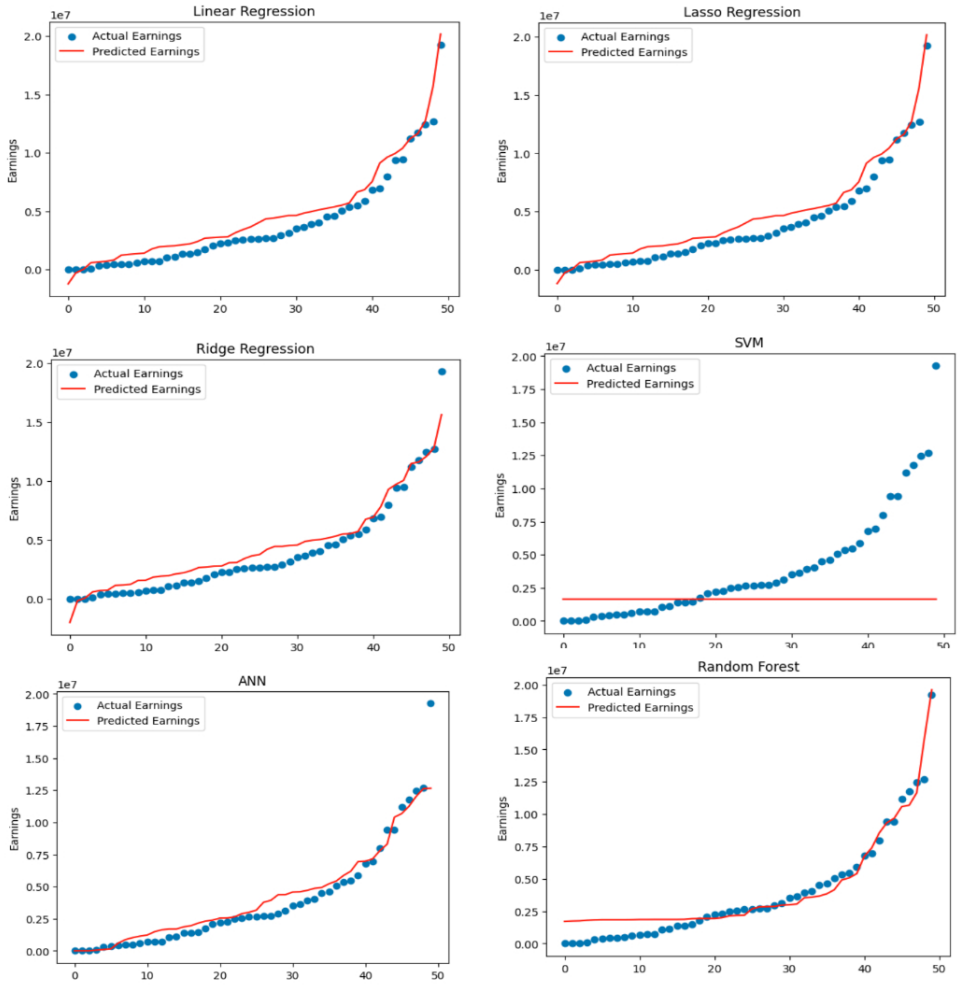
The following table compares the Machine Learning models and their results based on the evaluation of Root Mean Square Error (RMSE) and R- squared (R2) Error.

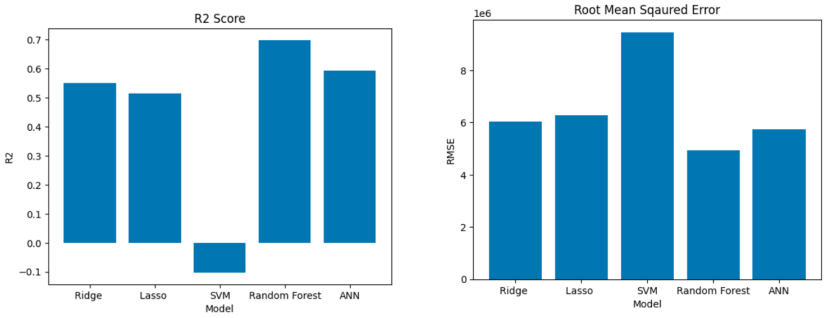
The evaluation of regression models, conducted through RMSE and R2 Error metrics, reflects substantial challenges attributed to the scale of the 'Earnings' target feature, typically in the range of $\$ 10^7$. The amplified impact of minor prediction errors due to this large scale emphasizes the sensitivity of RMSE to deviations in absolute terms. Even a slight miscalculation, such as an error of $1 million, significantly influences RMSE for earnings within this substantial range. Furthermore, squaring these discrepancies in RMSE calculations magnifies the differences between predicted and actual values, culminating in the observed higher RMSE values across models.

Notably, the occurrence of negative R2 values in Linear Regression and SVM Regression delineates an alarming inadequacy in capturing the intricate relationship between channel demographics and earnings. This highlights an urgent need for reassessing the model's architecture, refining feature selection methodologies, and scrutinizing dataset quality to augment predictive accuracy and overall model efficacy.

Regression Model	RMSE	R2 Error	Explanation
Linear Regression	2.357×10^{18}	-6.846×10^{22}	The extremely high RMSE and negative R2 Error suggest poor model fit or overfitting issues, yielding inaccurate predictions.
Lasso Regression	6048748.07	0.55	Exhibited moderate performance with a decent R2 score, indicating better fit and potential feature selection benefits from the penalty term.
Ridge Regression	6272516.38	0.51	Similar to Lasso, Ridge Regression displayed moderate performance with slightly higher RMSE and slightly lower R2, potentially handling multicollinearity better.
SVM Regression	9460147.02	-0.10	SVM Regression performed poorly with a high RMSE and low R2, possibly due to suboptimal kernel or hyperparameters.
Random Forest Regression	5130167.68	0.67	Random Forest Regression showcased better performance with moderate RMSE and higher R2, indicating robustness in handling nonlinear relationships.
ANN	5543106.18	0.62	The ANN model demonstrated moderate performance, with relatively lower RMSE and higher R2, suggesting its ability to capture complex patterns.

Table 1: Summary of regression model performance





(Note: The evaluation metrics for Linear Regression have been omitted due to poor fit and metric values that are off the scale.)

While Lasso and Ridge Regression models demonstrated moderate performance, leveraging regularization properties aiding in feature selection and handling multicollinearity, SVM Regression's subpar performance underscores the necessity for enhanced parameter optimization. Conversely, the superior performance of Random Forest Regression and ANN suggests their adeptness in capturing nonlinear relationships and intricate patterns within the dataset.

5 Conclusion:

Our project encompassed extensive data preprocessing techniques to predict YouTube channel earnings based on diverse demographics. Evaluation of machine learning models identified Random Forest Regression as most effective, closely trailed by ANN, demonstrating lower RMSE and higher R2. Conversely, Linear Regression fared poorly, displaying high RMSE and a negative R2, signifying its inadequate fit to the data. We discovered the significance of refined feature engineering and the potential of diverse regression techniques, emphasizing the need for sophisticated algorithms.

Additionally, while acknowledging the project's progress, time constraints restricted the exploration of exhaustive hyperparameter tuning in SVM Regression and novel feature engineering methods. These unexplored paths hold promise for future investigation, potentially enhancing model performance in predicting YouTube earnings.

6 Appendix:

Work Distribution:

Atharva Pansare: Data Preprocessing and Data Visualization, Training Ridge Regression, Lasso Regression and Support Vector Regression Models to predict Earnings and documentation.

Ritwik Jog: Literature Survey, Data Visualization, Implemeting PCA for dimensionality reduction, training Linear Regression and Random Forest models to predict earnings and documentation.

Viraj Sanap: Literature Survey, Data Visualization, Implementing Artificial Neural Network models to predict earnings and documentation.

Project GitHub Link: https://github.ncsu.edu/rsjog/engr-ALDA-Fall2023-P32