TEAM: DYNAMIC DUDES

Followers Gain Prediction of Top Streamers on Twitch

Team:

- Charan Tej K. (BL.EN.U4AIE190**13**)
- Chavali Surya Teja (BL.EN.U4AIE190**14)**
- Sugash T.M. (BL.EN.U4AIE190**62)**

Abstract

Our topic for the presentation is "**Followers gain prediction of top Twitch streamers**". In this project, we are going to train our machine learning model to predict the increment in followers of a particular streamer on the cloud data or a random streamer.

We are building our model using the **Multivariable Linear Regression** machine learning algorithm to teach the model.

Talking about the dataset, as the topic suggests, the whole data consists of relevant information about some top-notch individual twitch streamers such as Watch time (in minutes), Stream time (in minutes), Peak Viewers, Average Viewers, Number of Followers, Views Gained, Language, Partnership with Twitch (True/False), Mature Content in a stream (Yes/No). So, as we see the last two features can be formulated as binary outcomes 0 and 1.

The main input variables to the model would be the unique features which are the Watch time, Stream time, Peak Viewers, Average Viewers, Number of Followers, Views Gained and the output predicted variable would be the Followers gained.

TEAM: DYNAMIC DUDES

Data set explanation

The Data set we have used is the "**Top Streamers on Twitch Data set**", taken from the Kaggle website:

https://www.kaggle.com/aayushmishra1512/twitchdata

The data set mainly talks about individual descriptive features regarding all the top 1000 streamers on twitch:

- **Stream Time:** The watch time is the total time in minutes, that particular streamer streamed a game or a video live.

- **Watch Time:** The stream time column is the total time in minutes, the users or viewers watched a specific streamers video or a game.

- **Peak Viewers:** This feature speaks about the maximum number of viewers at a particular given instance for that distinct streamer.

- **Average Viewers:** Average viewer's feature is nothing but the average amount of viewers or users watching the streamers video on the whole.

- **Views Gained:** This column is about the total number of views gained for the streamer based on the number of users or watchers viewing that particular stream in total.

- **Number of Followers:** The number of followers is the total amount of users following that certain streamer.

- **Partnered with Twitch:** This is feature has only 2 values as outcomes and hence a binary feature. It is mainly about whether the selected streamer has partnered with Twitch or not.

- **Mature Content:** Similar to the above descriptive feature, this independent variable too, is a binary feature with Yes/No as the outcome. It talks about whether a particular stream contains matured content or not, i.e, 18+, or not.

- **Language:** This feature mainly contains the language of that particular stream (mostly being English).

Machine Learning Theory

As mentioned previously, we are using "**Multivariable Linear Regression**" to train the machine learning model to predict the target variable.

Reason behind choosing Multivariable Linear Regression:

As the basic point to be noted, our target is a continuous value, so it is a viable option to use Regression analysis, rather than Classification techniques.

Now,

since our input independent variables are more than one (six), and the predicted target variable is one, it is feasible to use the Multivariable Linear Regression algorithm rather than Simple Linear Regression

Algorithm, since Simple Linear Regression can only come into the picture when the input variable is only one.

Multivariable Linear Regression:

Multiple Linear Regression is one of the **Supervised** Machine Learning Statistical Techniques used to model the linear relationship between one continuous dependent (response) variable and two or more independent (explanatory) variables.

As far as our data set is concerned,

Target Variable – Followers Gained by the Streamer

Input Variables - Watch Time, Streaming Time, Peak Viewers, Average Viewers, Number of Followers, Views gained.

Model $M_{(w)}$ and Error Function:

Our basic formula is:

$H(\theta) = M_{(w)}(d) = w[0] + w[1] * d[1]$,

[This can be visualized as the simple slope-intercept form of a line,

$y = m*x + c$]

where 'w' is the vector (w [0], w [1]) and the parameters w [0] and w [1] are referred to as weights, 'd' is an instance defined by a single descriptive feature d [1], $M_w(d)$ is the predication output by the model for the instance d.

Now,

The above equation can be interpreted for Multivariable Linear Regression as follows:

TEAM: DYNAMIC DUDES

$H(\theta) = M_{(w)}(d) = w[0] + w[1] * d[1] + w[2] * d[2] + \ldots\ldots + w[m] * d[m]$,

for 'm' descriptive features.

Taking into account the above-generalized equation, we can modify the formula according to our data as follows:

Followers Gained = w [0] * d [0] + w [1] * [Watch time] + w [2] * [Stream Time] + w [3] * [Peak Viewers] + w [4] * [Average viewers] + w [5] * [Number of followers] + w [6] * [Views gained].

- To find the optimal set of weights, we need an optimal measure to tell us how well a model is defined using a set of weights that fit a training dataset.

  For that, we bring in:

  <u>The sum of squared errors error function L_2:</u>

  An error function helps us calculate the error between the predictions made by a model and the actual values in a training dataset.

$$\textbf{J(w[0], w[1], w[2], w[3], w[4], w[5], w[6])} = \boldsymbol{L_2(M_w, d)} = \left(\frac{1}{2}\right)\sum_{j=1}^{m}(t_j - (w * d_j))^2.$$

The above equation is modeled using the **Vector L_2 Norm**, which is one of the three basic Vector Norms for machine Regression analysis.

TEAM: DYNAMIC DUDES

Implementation Steps

- The very first step in implementing the Machine Learning Algorithm on our data would be to import all the important and necessary dependencies.

- We then move to visualize the data using the imported libraries (mentioned above) such as **Matplotlib**, **Seaborn**, etc., to get a clear picture of our data set and to very well understand the behavior of the input independent explanatory variables on the predicted target variable.

- Moving on to modeling the data, we import required libraries and internal modules for scientific computing such as **sklearn** which contains modules namely, **LinearRegression**, **StandardScaler()**, **train_test_split**, etc..

- We train our data using the split ratio of 80:20 is as to the Training data: Testing data. (Since it is left over to the user's freedom, we can alter the values of the above split ratio).

- Finally, we predict the target variable which is Followers Gained, using the six explanatory variables as the input, which are – Watch Time, Stream time, Peak Viewers, Average Viewers, Number of Followers, Views Gained.

- At last, we visualize all sets predicted values by plotting the best fit line using the seaborn's **Regplot** module.

TEAM: DYNAMIC DUDES

Sample Code and Output

Code for predicting the Followers Gained:

```
user_input = [[744620970,118125,26141,6328,859439,15894732]]
user_pred = linear.predict(user_input)
print("Follower Gained by the streamer are:-",user_pred)
```

Output:

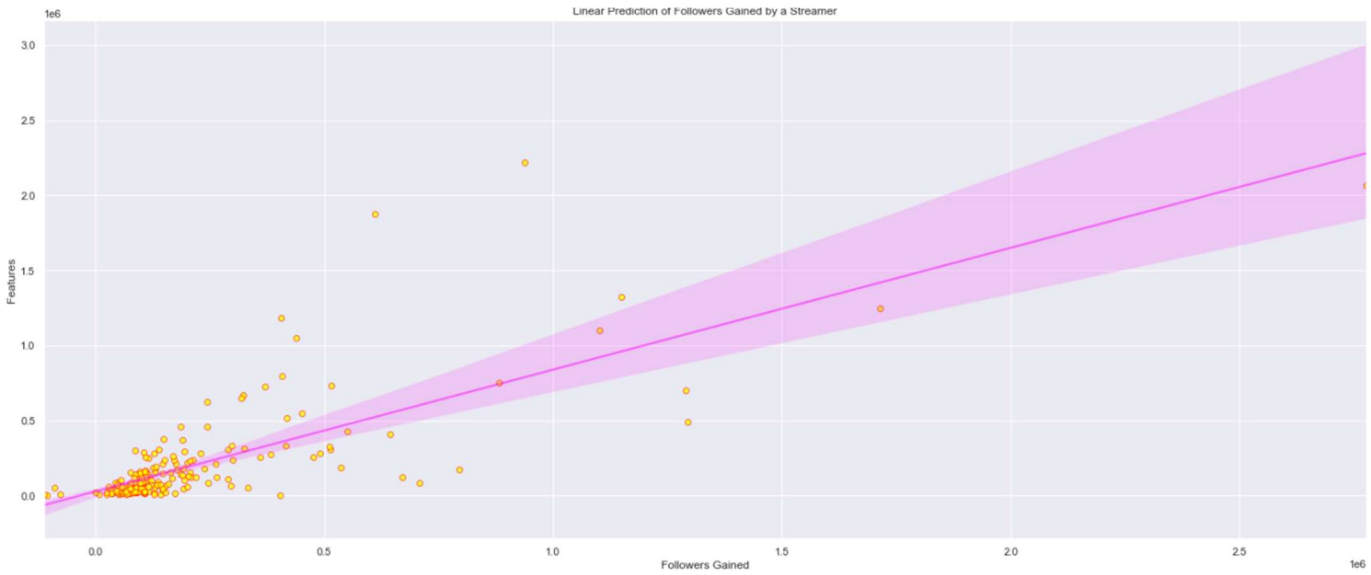
```
Follower Gained by the streamer are:- [3.3585976e+13]
```

Code for plotting the best fit line:

```
plt.style.use('seaborn-dark-palette')
plt.figure(figsize=(25,10))
sns.regplot(pred,y_test,scatter_kws={'color':'yellow','edgecolor':'red','linewidth':0.7},line_kws=
{'color':'magenta','alpha':0.5})
plt.xlabel('Followers Gained')
plt.ylabel('Features')
plt.title("Linear Prediction of Followers Gained by a Streamer")
plt.show()
```

TEAM: DYNAMIC DUDES

Output:



Conclusion:

From the regression plot, we could infer that most of the followers gained lies between **$0.1x10^6$ – $0.8x10^6$**. We can also come to the end that each feature showed a different effect on the target variable at different times. We can observe that Multivariable Linear Regression has a great advantage over Simple Linear Regression.

Finally, I conclude by saying that we as a whole team understood the importance of Machine Learning Algorithms, their implementation in a programming language, and the mathematics involved in them.

WRITTEN BY: SURYA TEJA CHAVALI

----------*****----------