Predicting top 1000 Youtube Channels Average Yearly Earning

Abstract

The purpose of this project is to scrape Popsonner and SocialBlade to see what might affect Average yearly earning for the top 1000 youtube channels and predict the Average yearly earning

Design

We scraped two websites(Popsonner and SocialBlade)and combined features to make the model more reliable.

Data

First website which called: Popsonner, we scrapped the following features: Channels IDs, Subscriber, Views, Position, Published and Videos. Second website is SocialBlad, we got more features such as: Social Blade Rank, Country Rank, Video Views Rank, Channel Type, Music Rank, Estimated monthly Earning, Subscribe for the last 30 days, Video viwes for the last 30 days

Analysis Steps

1- Scrapping the data from both websites, and merge them on Channel ID. 2- Create a dataframe and start performing EDA: + Dropping nulls, removing outliers, removing duplicates 3- Applying improvement features: + feature selecting: to select important features that help in predictting our target + feature engineering: 1- creating Average yearly earning and Average Monthly Earning out of min/max estimated monthly earning 2- Encoding categorical variables: Encode with value between 0 and n_classes -1 4- Test Linear regression assumptions:test linearity between dependent and independent variables 5- Applying log transformation to independent numerical variables to increase the correlation 6- Fitting linear regression model 7- improve the model by fitting polynomial to different degrees and Ridge and Lasso 8- make a decision about models based on (R Squared and MSE)

Conclusion and summary

Eventhough polynomial showed low MSE, R squared is high to level that we reach overfitting. Thus, linear regression isn't the appropriate model to meet our goal!