أكاديمية سدايا
SDAIA Academy

METIS

PREDICTING THE AVERAGE YEARLY EARNING

FOR THE TOP 1000 CHANNELS ON YOUTUBE

By : Manal AlQahtani, Nadia AlGhamdi

As Project 2 of SDAIA Data Science Bootcamp (T5)

أكاديمية سدايا
SDAIA Academy
METIS

OUTLINE

01 INTRODUCTION

02 WORKFLOW

03 TOOLS

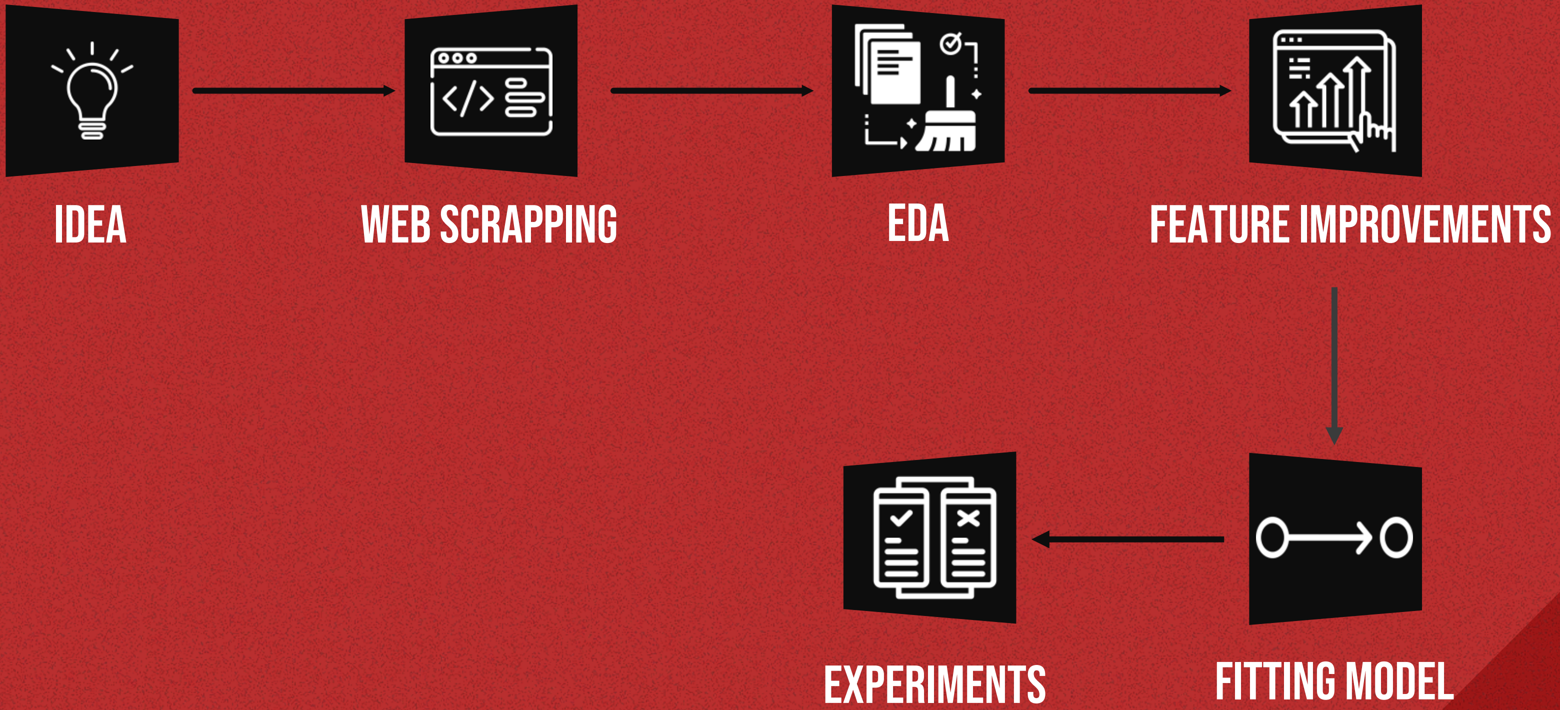
04 IMPLEMENTATION

05 CONCLUSION

INTRODUCTION

- Nowadays Social media is considered a wealth source, where everyone can make profit out of it .
- And the YouTube is considered half the internet, where 1.9 billion users logging in it .

WORKFLOW


IDEA
WEB SCRAPPING
EDA
FEATURE IMPROVEMENTS
FITTING MODEL
EXPERIMENTS

أكاديمية سدايا
SDAIA Academy
METIS

TOOLS

PYTHON

Jupyter notebook	BeautifulSoup
NumPy, Pandas	Selenium
Matplotlib, Seaborn	Sklearn

HTML

- HTML
- CSS

أكاديمية سدايا
SDAIA Academy
METIS

WEB SCRAPING

SCRAPING

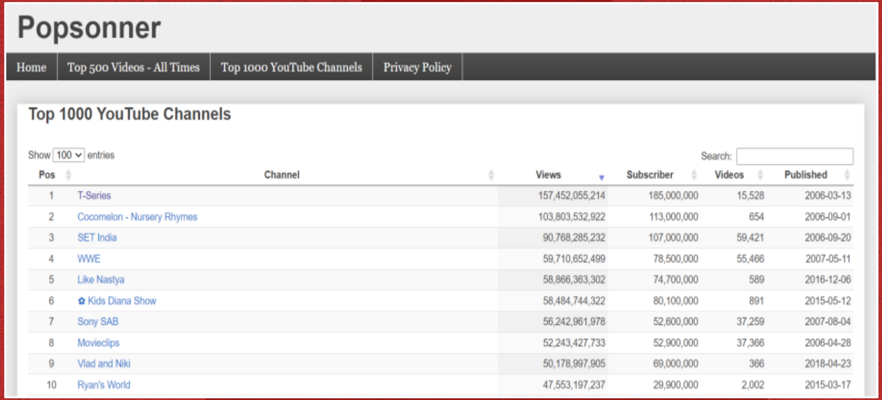
- Popsonner
- Social Blade

+ 1K ROWS

15 FEATURES

أكاديمية سدايا
SDAIA Academy
METIS

DATASET

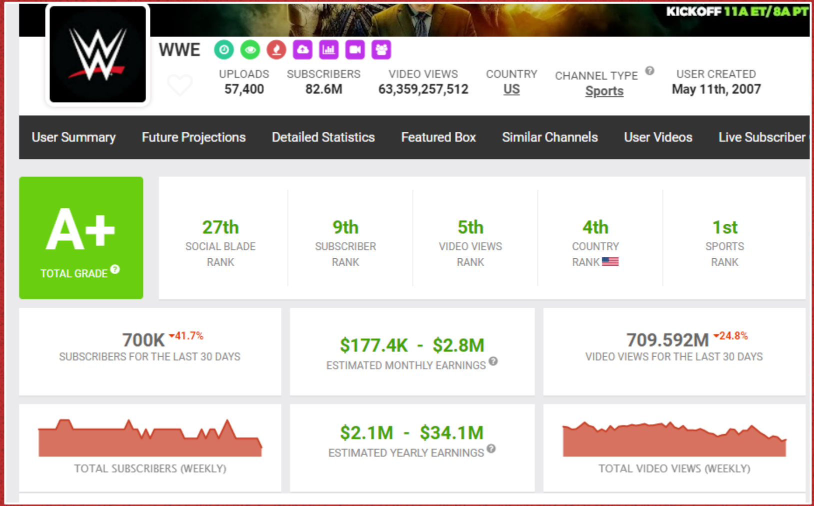

Popsonner

Home | Top 500 Videos - All Times | Top 1000 YouTube Channels | Privacy Policy

Top 1000 YouTube Channels

Show 100 entries | Search:

Pos	Channel	Views	Subscriber	Videos	Published
1	T-Series	157,452,055,214	185,000,000	15,528	2006-03-13
2	Cocomelon - Nursery Rhymes	103,803,532,922	113,000,000	654	2006-09-01
3	SET India	90,768,285,232	107,000,000	59,421	2006-09-20
4	WWE	59,710,652,499	78,500,000	55,466	2007-05-11
5	Like Nastya	58,866,363,302	74,700,000	589	2016-12-06
6	✿ Kids Diana Show	58,484,744,322	80,100,000	891	2015-05-12
7	Sony SAB	56,242,961,978	52,600,000	37,259	2007-08-04
8	Movieclips	52,243,427,733	52,900,000	37,366	2006-04-28
9	Vlad and Niki	50,178,997,905	69,000,000	366	2018-04-23
10	Ryan's World	47,553,197,237	29,900,000	2,002	2015-03-17



- Merged 2 datasets on Channel ID
- Create DataFrame out of them

EDA

NULLS VALUES

CHANGE COLUMN TYPES

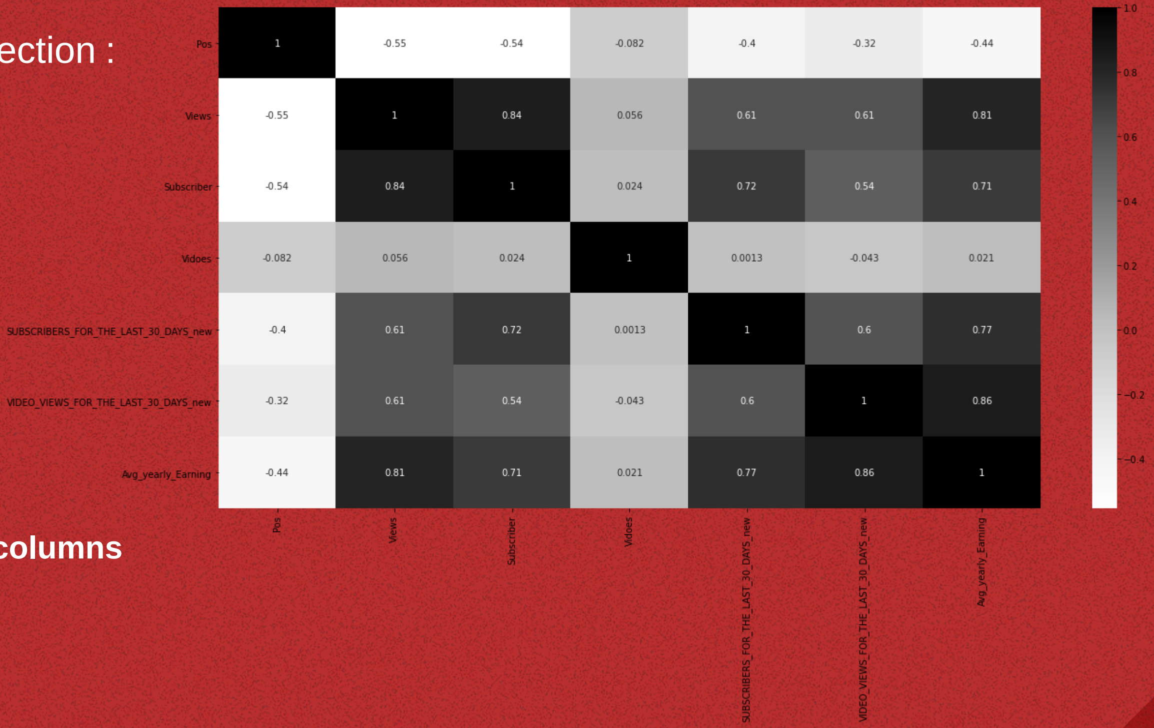
OUTLIERS

أكاديمية سدايا
SDAIA Academy
METIS

EDA

- Features Engineering :
    Getting the Average monthly/ yearly earning out of min and max .

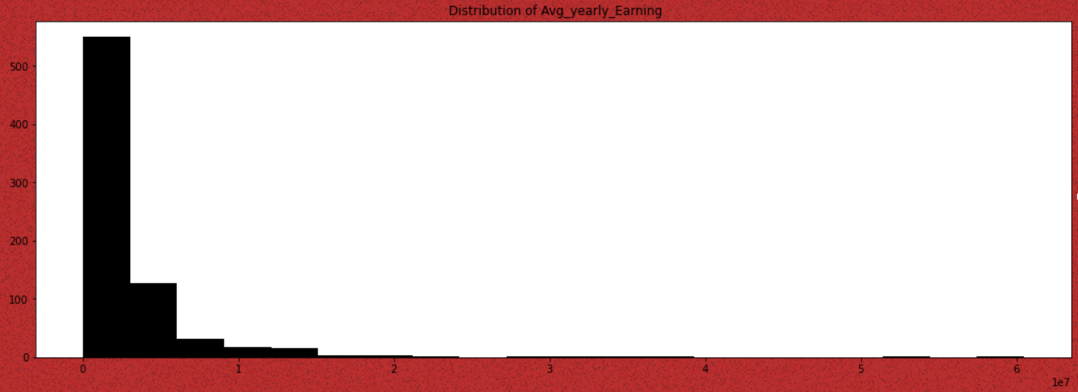
- Features Selection :

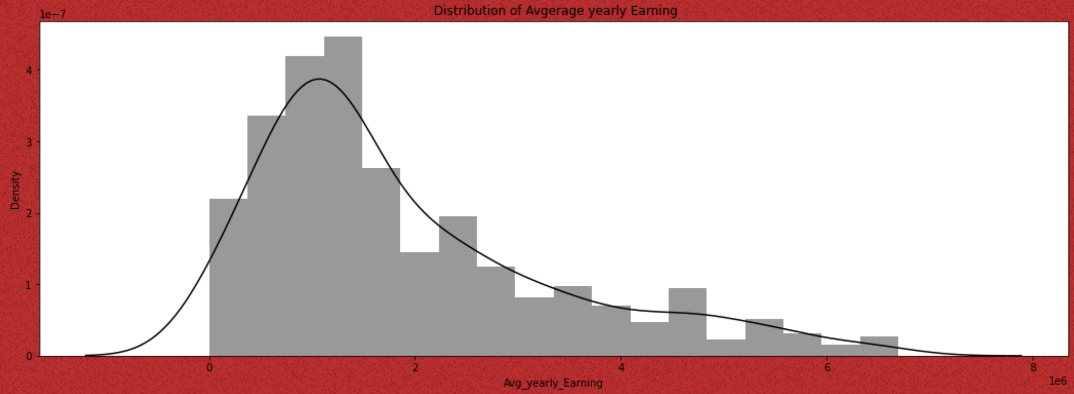


687 rows × 13 columns

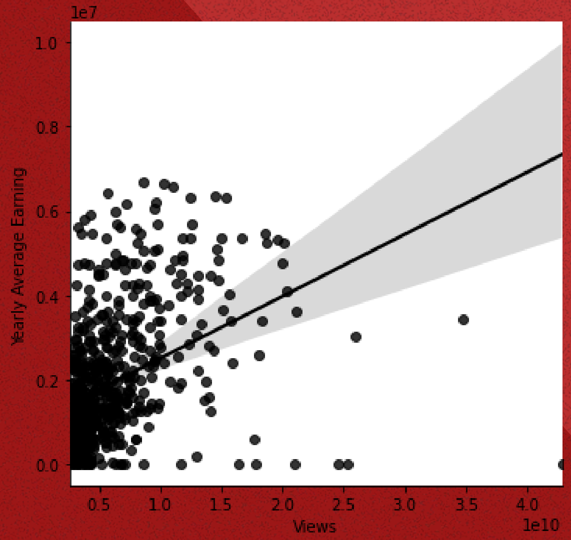
EDA

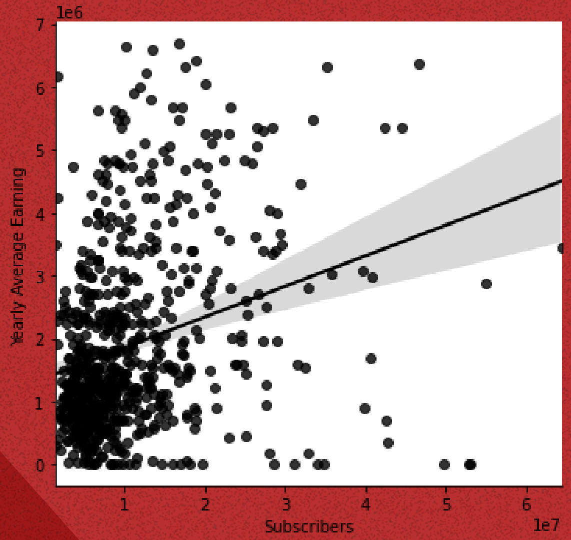
- Outliers :

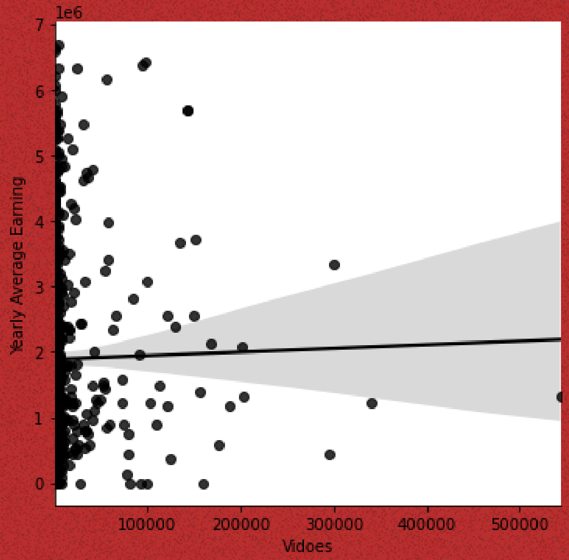




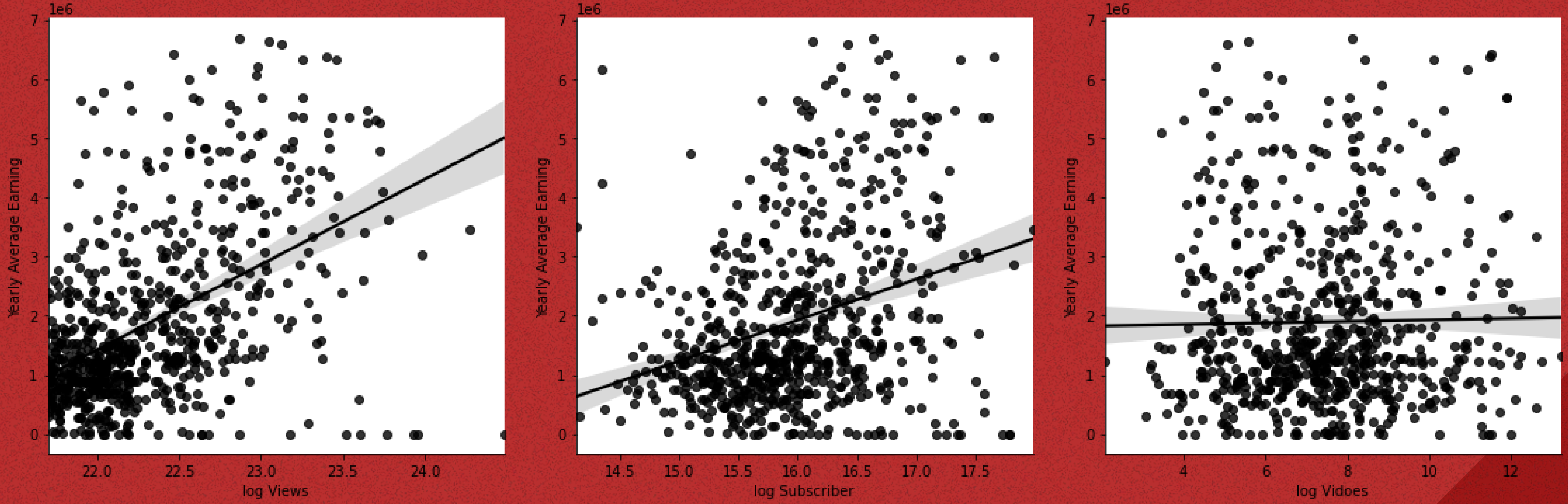
- Testing Linearity between dependent and independents variables :







- Log Transfromation to establish linear correlation between Y and features :



أكاديمية سدايا
SDAIA Academy
METIS

- Encoding categorical variables :

encode with value between 0 and n_classes – 1 .

أكاديمية سدايا
SDAIA Academy
METIS

MODELING

- Linear regression result :

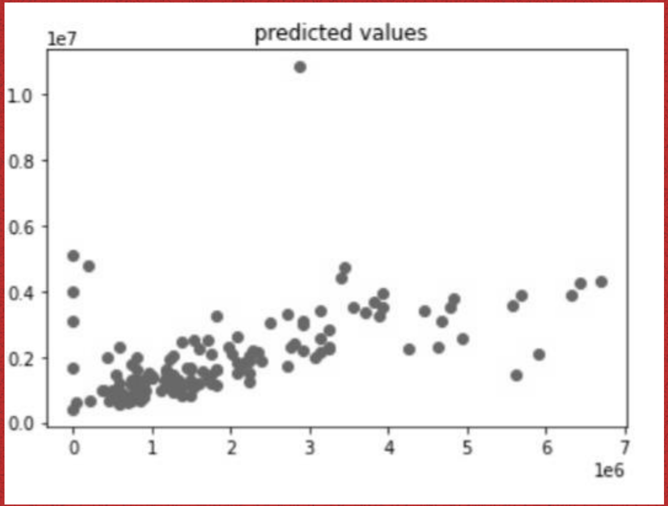
	r_squared	MAE	MSE	RMSE
Linear Regression	0.634396	653861.543143	8.949487e+11	9.460173e+05

```
Index(['Pos', 'Channel', 'SOCIAL_BLADE_RANK', 'VIDEO_VIEWS_RANK',
       'COUNTRY_RANK', 'MUSIC_RANK', 'CHANNEL_TYPE',
       'SUBSCRIBERS_FOR_THE_LAST_30_DAYS_new',
       'VIDEO_VIEWS_FOR_THE_LAST_30_DAYS_new', 'Avg_yearly_Earning', 'Views',
       'Subscriber', 'Vidoes'],
      dtype='object')
```

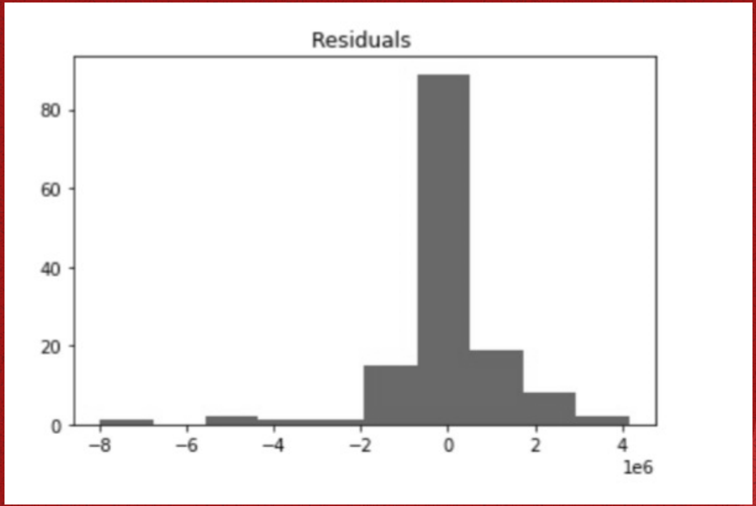
```
r = pd.DataFrame(lr.coef_, columns = ['Coeffcients'])
r
```

	Coeffcients
0	1.574779e+03
1	-1.184288e+02
2	-6.511618e+01
3	-1.164976e+02
4	-4.005173e+02
5	-6.594658e+02
6	2.075652e+04
7	6.866476e+00
8	3.876942e-03
9	2.492482e+06
10	-5.830345e+05
11	-2.199701e+03

- Making prediction from LR



- Assess the performance of our model :



- **Improving Linear Regression**

	r_squared	MAE	MSE	RMSE
Linear Regression	0.634396	653861.543143	8.949487e+11	9.460173e+05
Polynomial degree 4	0.951278	221565.225829	1.192647e+11	3.453473e+05
Polynomial degree 5	0.299057	787351.038821	1.715812e+12	1.309890e+06

- Regulization :

	R Squared	MAE	MSE	RMSE	Best Alpha
Lasso	0.471861	571276.702456	9.130710e+11	955547.479501	0.0
Ridge	0.471861	571276.702456	9.130710e+11	955547.479501	0.0
Elastic Net	0.471861	571276.702456	9.130710e+11	955547.479501	0.0

CONCLUSION

Even though 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 !

أكاديمية سدايا
SDAIA Academy

METIS

THANK YOU !

FOR YOUR KIND ATTENTION