Prediction of Expected Lifespan of a Satellite

1. PROBLEM SETTING

According to United Nations Office for Outer Space Affairs(UNOOA), there are about 4987 satellites orbiting the planet at the start of the year with the increment of 2.68% every year. However, there are only about 1957 satellites active which indicates that there's a lot flying metal lying out in the free space. We are intrigued on what happens to these satellites that are no longer in use and would like to further investigate and to avoid spacecraft collisions, predicting life span of the satellites is necessary.

2. PROBLEM DEFINITION:

The project aims at identifying the best suitable model for predicting the expected lifespan of a satellite. The main intention of the analysis is to understand the behavior of each variable and find their significance with the response variable using feature selection and how well they perform when fit into different models.

3. DATA SOURCE

- https://www.kaggle.com/ucsusa/active-satellites

  We obtained this dataset from Kaggle that includes all the details of satellites.

- https://www.euspaceimaging.com/the-lifespan-of-orbiting-satellites/

  The satellite information can be found the above websitE

4. DATA DESCRIPTION

The clean dataset has about 1500 records with 25 variables. There are 18 variables in object type and 6 variables in numeric and 1 variable in date-time format. Some of the most notable variables are described below.

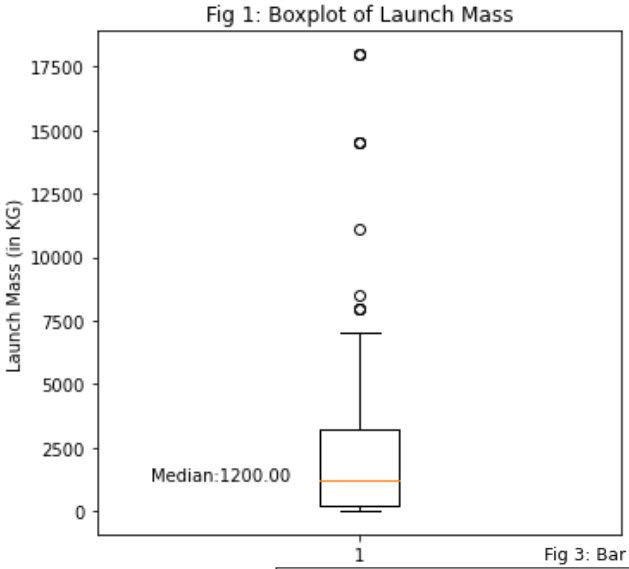
Variable	Definition
Official Name of Satellite	Name given to the satellite at launch
Country of UN Registry	Country sending the satellite
Users	Variable with 4 categories Military, Civil, Government, Commercial
Purpose	Reason for the launch of satellite
Launch Mass(kg)	Mass of the satellite during launch
Inclination(deg)	Tilt of a satellite orbit around earth
Power(watts)	Initial power of a satellite
Apogee(km)	Farthest point of satellite from earth
Perigee(km)	Closest point of satellite from earth
Eccentricity	Amount by which satellite orbit deviates from a perfect circle
Period (minutes)	Time taken by satellite to complete one orbit
Country	Country which has sent the satellite
Operator	Organization operating the satellite
Launch Date	Date of launch of satellite
Expected Lifespan (years)	Expected Lifespan of satellite

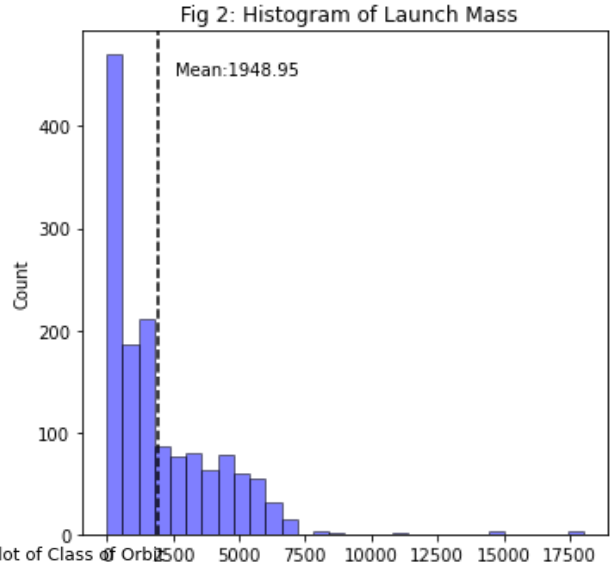
5. DATA EXPLORATION

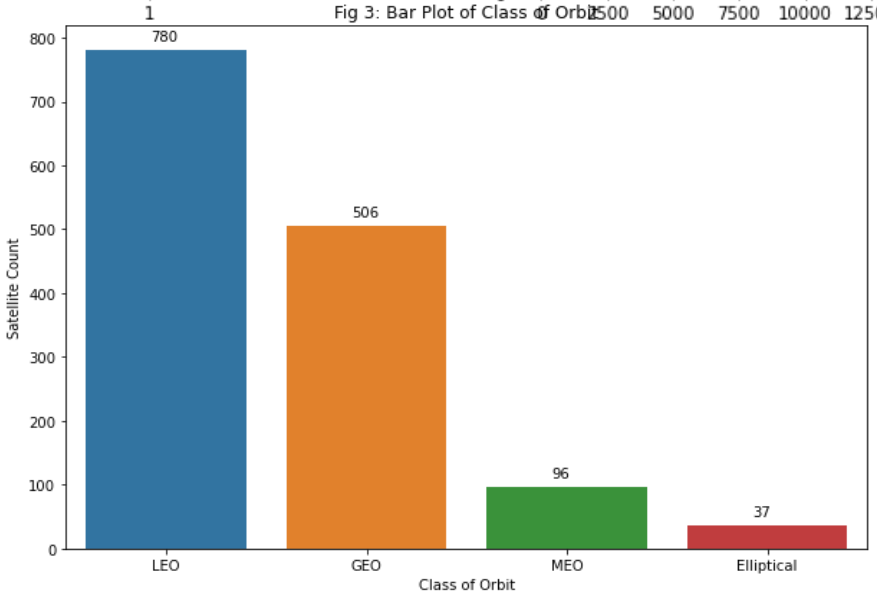
UNIVARIATE ANALYSIS:

For this section of our project, we examined the numerical and categorical variables separately using exploratory data analysis. The observations and results are listed below: We started off our exploration with univariate analysis of various both, categorical and numerical variables.

To understand the distribution of launch mass, we created a box plot and a histogram to find the range for the mass of satellites during the launch and it is evident that most of the satellites have a range of 250 - 3750Kg with an average launch mass of 1950 Kg(Fig 1,2)







We also visualized satellites orbiting in different class of orbits and we conclude that more than 50% of the satellites fall in the low earth orbital i.e having an altitude lower than 1000km.(Fig 3)

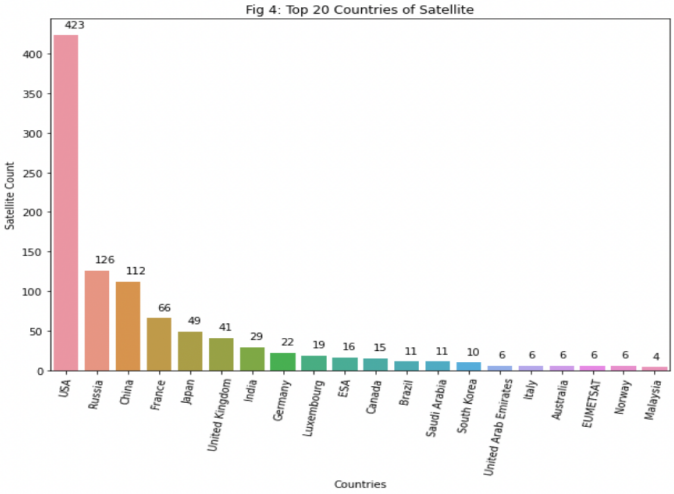


Fig 4: Top 20 Countries of Satellite

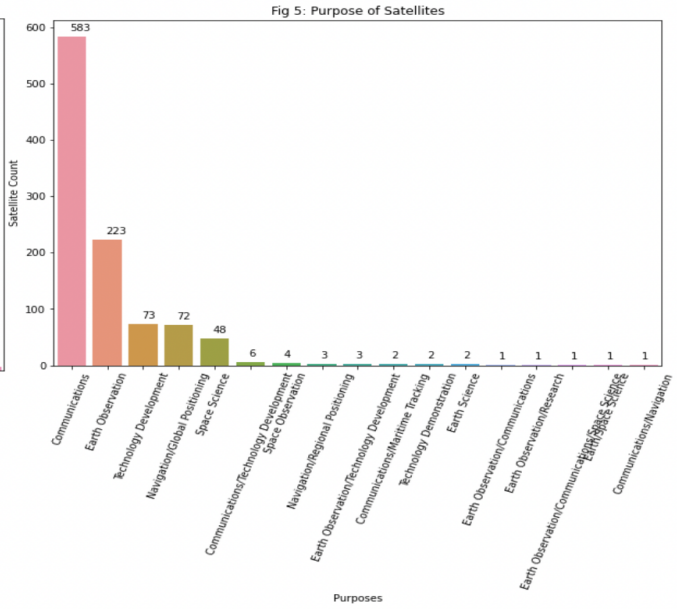


Fig 5: Purpose of Satellites

Fig 4,5 shows the count of top 20 countries with the most number of satellites and the count of top 20 purposes for using satellites. USA ranks the chart with the highest number of satellites with India and China standing in the top 10. While there are many purposes for sending satellites, Communications and Earth observatory satellites are the most used.

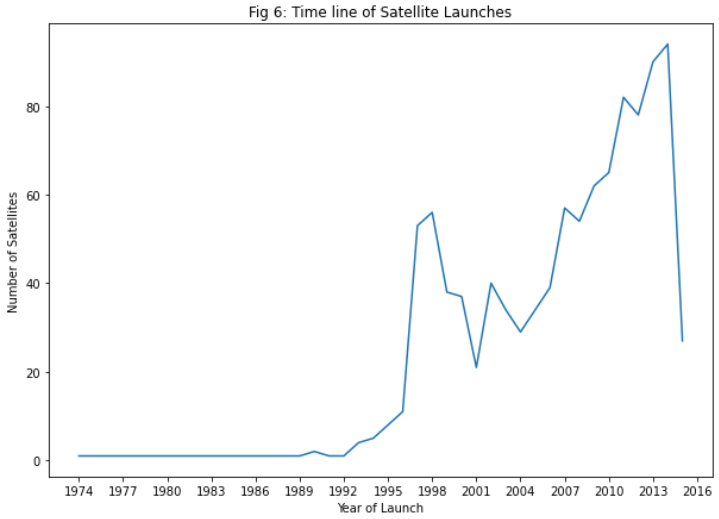


Fig 6: Time line of Satellite Launches

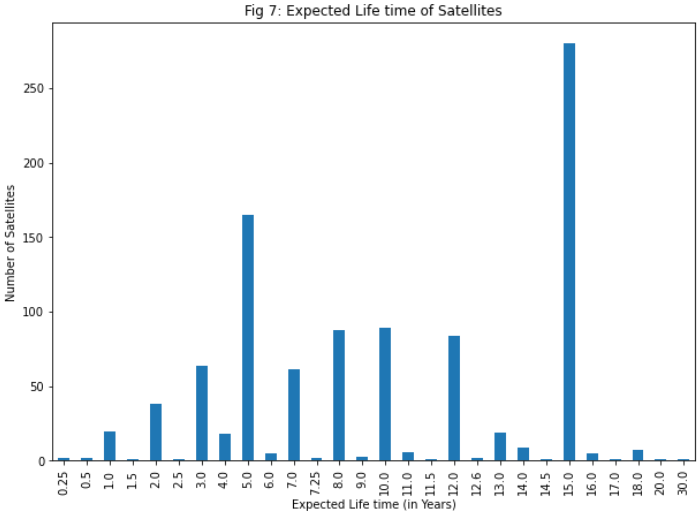
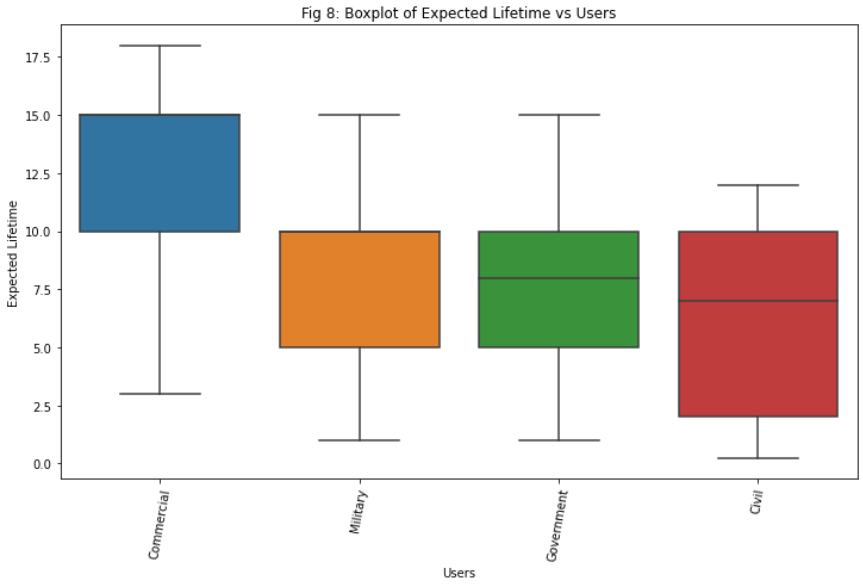


Fig 7: Expected Life time of Satellites

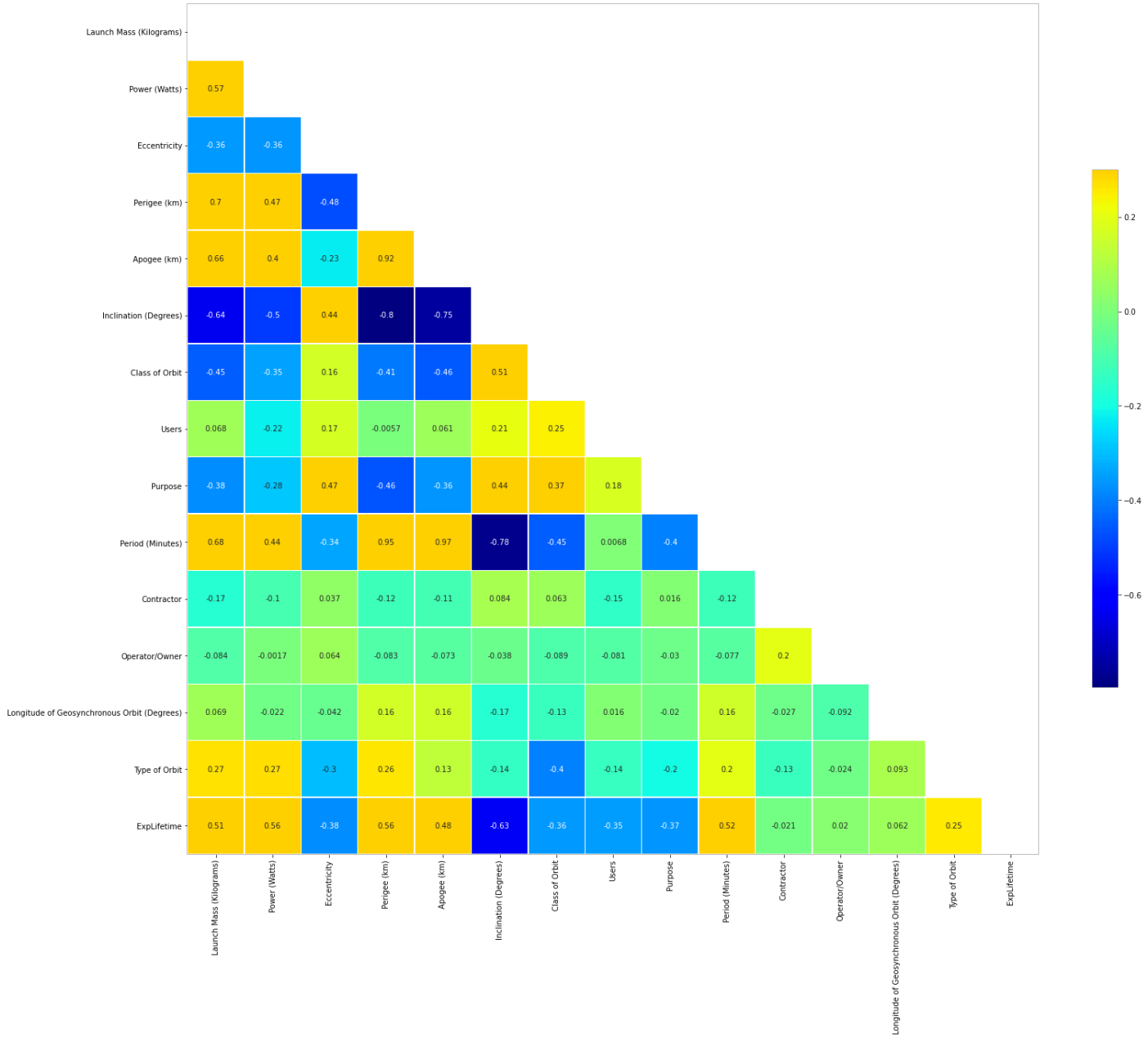
A line plot visualizing the number of satellites launched between the years 1974 to 2015 is shown below(Fig 6). The increased usage of satellites started in 1995 with a peak in 1998 which is evident from the plot. Since then, there is a gradual increase in the usage and number of the satellites launched. Another bar plot(Fig 7) visualizes the expected life time of a satellites and most of the satellites are expected to have a span between 5-15 years. A life span less than 1 year and more than 15 years is hardly observed.

BIVARIATE ANALYSIS:

We started the Bi-variate analysis with the distribution of different types of users of satellites with their expected lifetime. The most common users being commercial , civil, military and government. The expected lifetime of commercial satellites seems to be higher than the rest of the users from the boxplot below.



We used a heat map to find the significant co-relation between various variables and we observe that some variables have high correlation values. For example, apogee and perigee have strong correlation with a 0.91 coefficient and apogee-period with a 0.96 correlation coefficient while apogee and inclination exhibit a negative correlation with -0.83 correlation coefficient.


Launch Mass (Kilograms)
Power (Watts) 0.57
Eccentricity -0.36 -0.36
Perigee (km) 0.7 0.47 -0.48
Apogee (km) 0.66 0.4 -0.23 0.92
Inclination (Degrees) -0.64 -0.5 0.44 -0.8 -0.75
Class of Orbit -0.45 -0.35 0.16 -0.41 -0.46 0.51
Users 0.068 -0.22 0.17 -0.0057 0.061 0.21 0.25
Purpose -0.38 -0.28 0.47 -0.46 -0.36 0.44 0.37 0.18
Period (Minutes) 0.68 0.44 -0.34 0.95 0.97 -0.78 -0.45 0.0068 -0.4
Contractor -0.17 -0.1 0.037 -0.12 -0.11 0.084 0.063 -0.15 0.016 -0.12
Operator/Owner -0.084 -0.0017 0.064 -0.083 -0.073 -0.038 -0.089 -0.081 -0.03 -0.077 0.2
Longitude of Geosynchronous Orbit (Degrees) 0.069 -0.022 -0.042 0.16 0.16 -0.17 -0.13 0.016 -0.02 0.16 -0.027 -0.092
Type of Orbit 0.27 0.27 -0.3 0.26 0.13 -0.14 -0.4 -0.14 -0.2 0.2 -0.13 -0.024 0.093
ExpLifetime 0.51 0.56 -0.38 0.56 0.48 -0.63 -0.36 -0.35 -0.37 0.52 -0.021 0.02 0.062 0.25
Launch Mass (Kilograms)
Power (Watts)
Eccentricity
Perigee (km)
Apogee (km)
Inclination (Degrees)
Class of Orbit
Users
Purpose
Period (Minutes)
Contractor
Operator/Owner
Longitude of Geosynchronous Orbit (Degrees)
Type of Orbit
ExpLifetime
0.2
0.0
−0.2
−0.4
−0.6

6. FEATURE SELECTION:

For feature selection we have used Exhaustive Search method to obtain the best features. 'f_regression' is used as a scoring function, employs F-statistic which is given by:

$$PF_t = \frac{SSE_{t-1} - SSE_t}{MSE_{t-1}}$$

where,

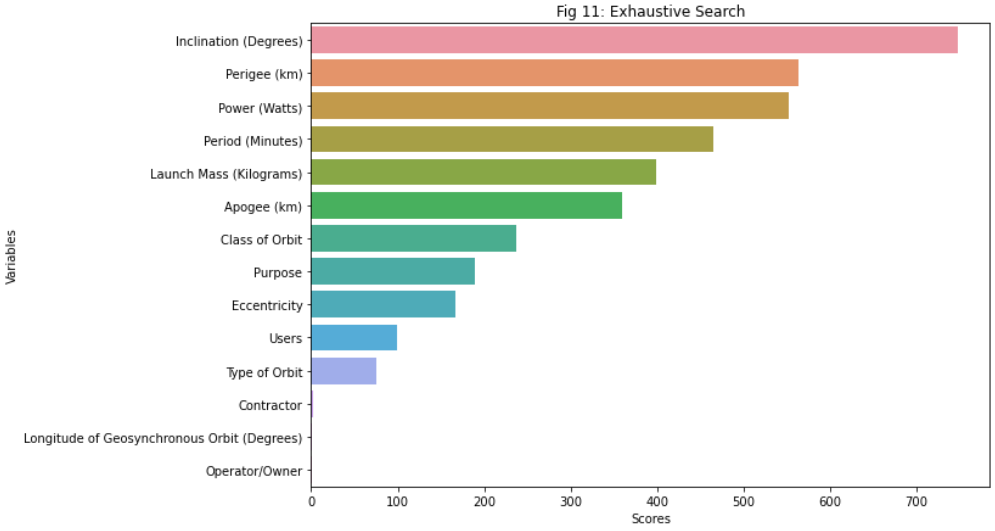
SSE_{t-1} = Residual sum of squares with t-1 predictors in model

SSE_t = Residual sum of square with t predictors when the additional t-th predictor enters in model

MSE_{t-1} = Mean squared error due to residuals with t -1 predictors

From the results obtained top 9 features were selected.

Variables	Scores
Inclination (Degrees)	747.399578
Perigee (km)	563.732084
Power (Watts)	552.685055
Period (Minutes)	465.237339
Launch Mass (Kilograms)	399.426115
Apogee (km)	359.529523
Class of Orbit	237.492062
Purpose	189.434530
Eccentricity	166.174446
Users	99.300532
Type of Orbit	75.324455
Contractor	2.218702
Longitude of Geosynchronous Orbit (Degrees)	1.015381
Operator/Owner	0.458218



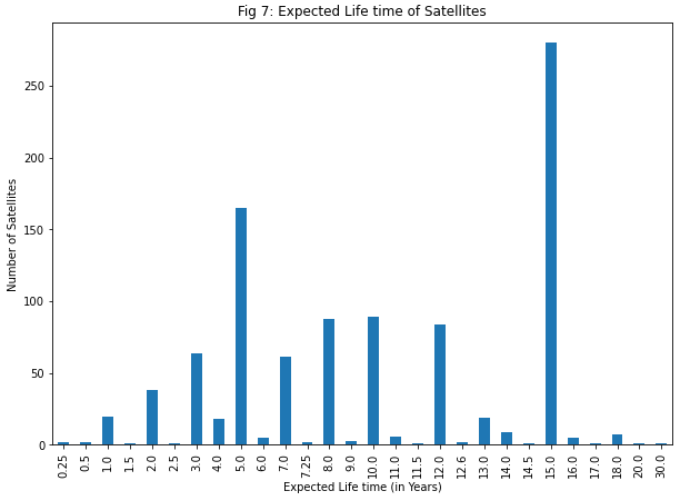
7. MODEL BUILDING:

The top 9 features from exhaustive search were selected as predictors -Inclination, Perigee, Power, Period, Launch Mass, Purpose, Apogee, Class of Orbit and Eccentricity and Expected Life Span is selected as a response variable.

The dataset is split into training and testing data with a test size 20%. After the split the train data consists 820 records and test data consists 206 records.

- **MODELING – MACHINE LEARNING:**

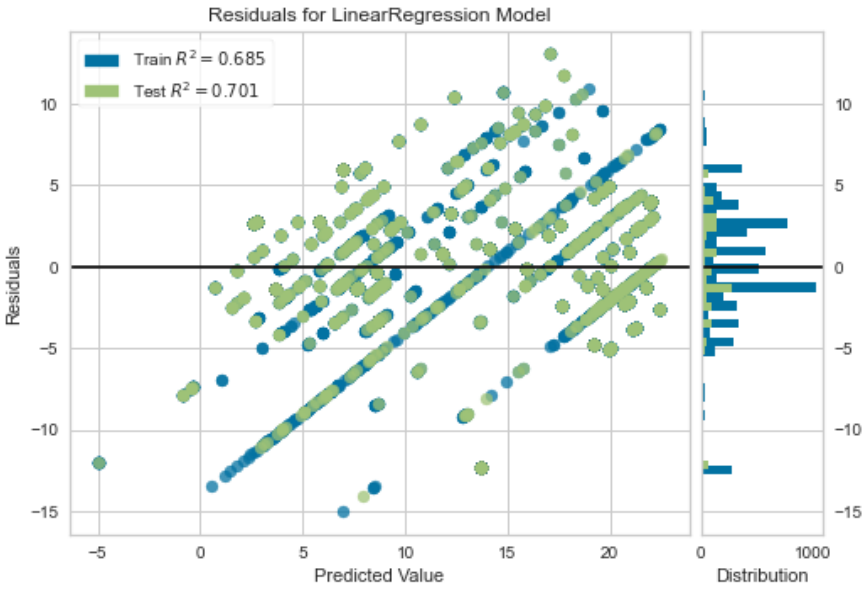
Most of the satellites have Expected Lifespan of 15 and 5 years, hence the data is imbalanced. So we have used Random Over Sampler to oversample the data before fitting.



Since our response variable is numeric we used mutliple regression models to fit our data. Some of the regression models used and their residual plots are as follows:

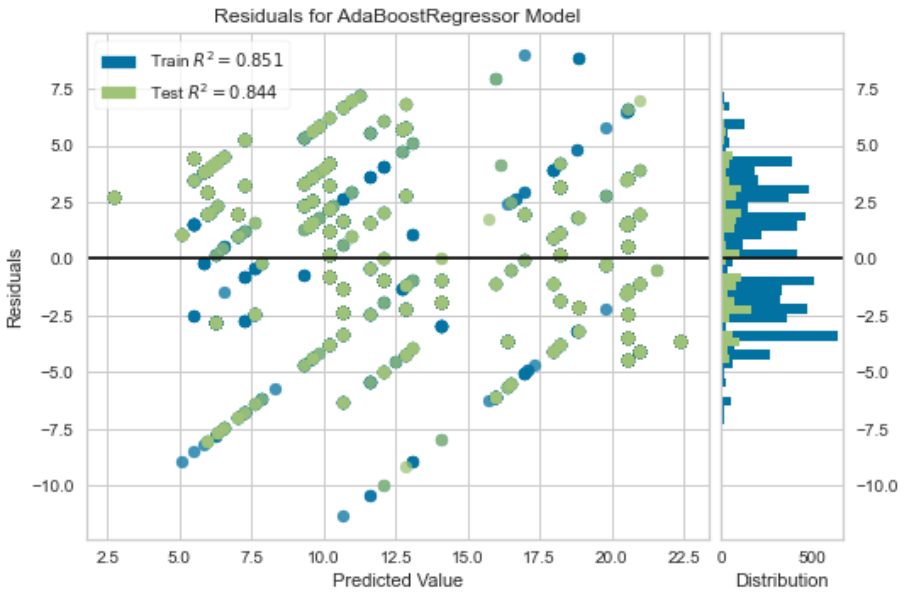
1) **Linear Regression:**

RMSE : 4.24



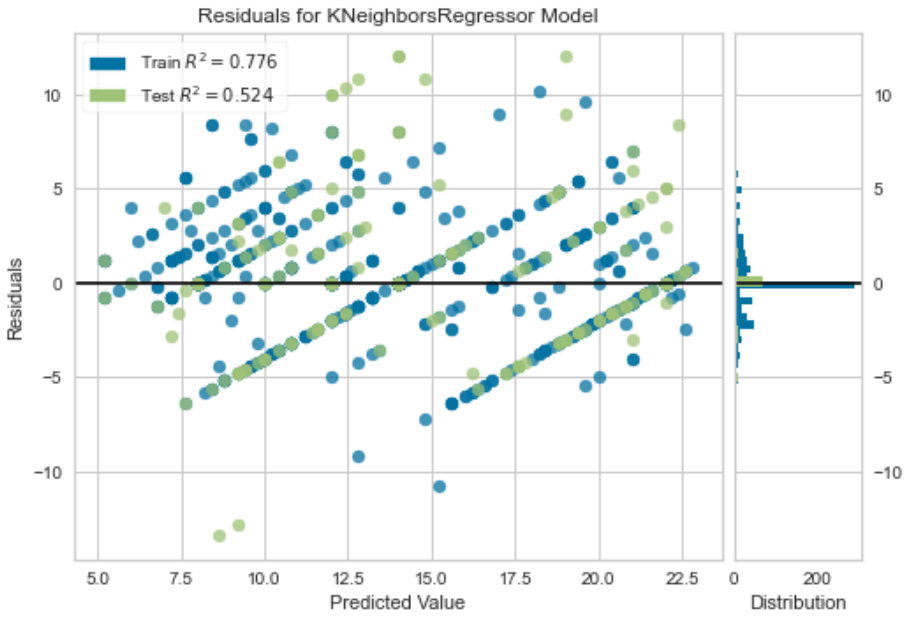
2) **Ada Boost Regression:**

RMSE : 3.065



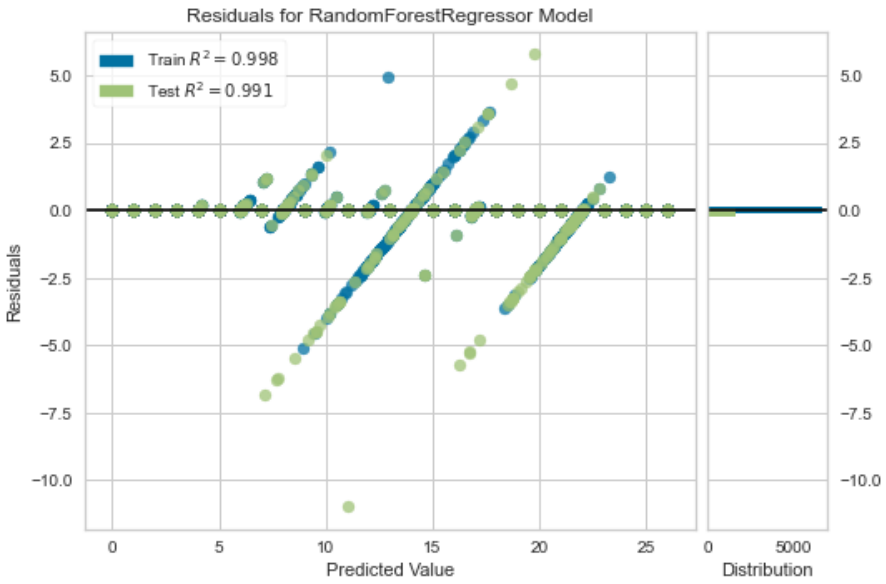
3) KNN Regression:

RMSE : 3.88



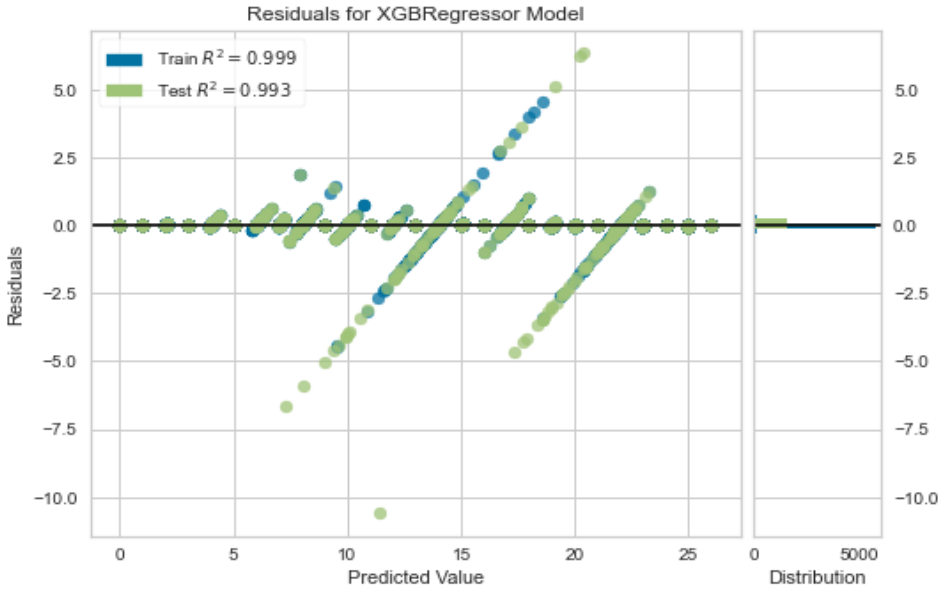
4) Random Forest Regression:

RMSE: 0.719



5) Xgboost Regression:

RMSE : 0.65



The residual plots of all the models are normal and the assumptions are not violated.

- **HYPERPARAMETER TUNING:**

In order improve the performance hyperparameter tuning of XGBoost Regressor is done using Randomized Search CV.

The parameter values given to Randomized search CV are:

```
#the rate at which the model learns
learning_rate = [0.05, 0.1, 0.15, 0.2, 0.25, 0.3]
#maximum number of levels in a tree
max_depth = [3,4,5,6,8,10,12,15]
# Minimum loss reduction required to make a further partition on a leaf node of the tree
gamma = [0.0, 0.1, 0.2, 0.3, 0.4, 0.5]
#Minimum sum of instance weight(hessian) needed in a child.
min_child_weight = [1,3,5,7,9]
#Subsample ratio of columns when constructing each tree
colsample_bytree = [0.3,0.4,0.5,0.7,0.9]
```

The best parameters obtained are:

```
XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
             colsample_bynode=1, colsample_bytree=0.7, gamma=0.5, gpu_id=-1,
             importance_type='gain', interaction_constraints='',
             learning_rate=0.25, max_delta_step=0, max_depth=6,
             min_child_weight=3, missing=nan, monotone_constraints='()',
             n_estimators=100, n_jobs=8, num_parallel_tree=1, random_state=0,
             reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1,
             tree_method='exact', validate_parameters=1, verbosity=None)
```

After 5 iterationd and 10 fold cross validation the RMSE after hyperparameter tuning reduced from 0.65 to 0.57 and hence the performance was boosted.

- **MODELING – DEEP LEARNING:**

We constructed a deep neural network using tensorflow with 9 nodes in input layer, 2 hidden layers with 9 nodes in each layer and an output layer.

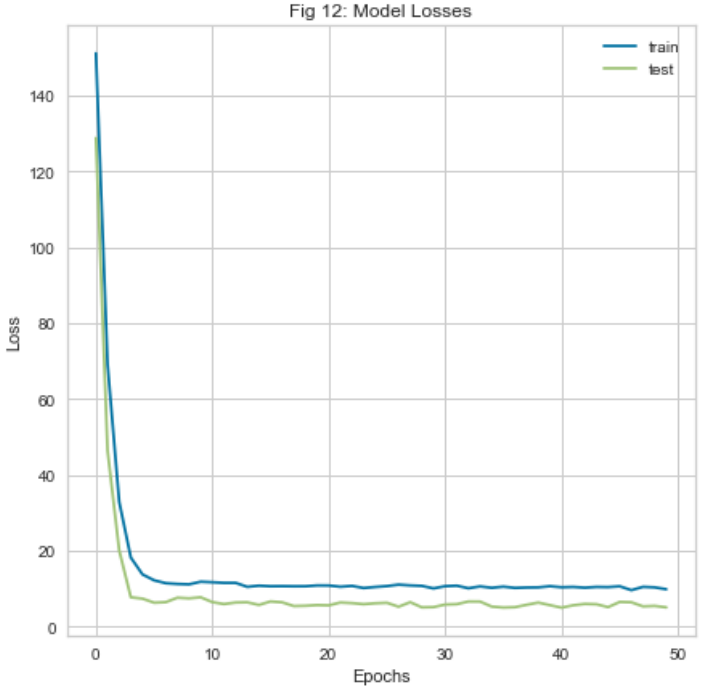
We used Leaky Relu as activation function for all layers and Adam optimizer learning rate 0.001.

After training the model for 50 epochs and batch size of 16, the results found were impressive.

RMSE : 1.57

R2 Score : 0.91

Fig 12: Model Losses

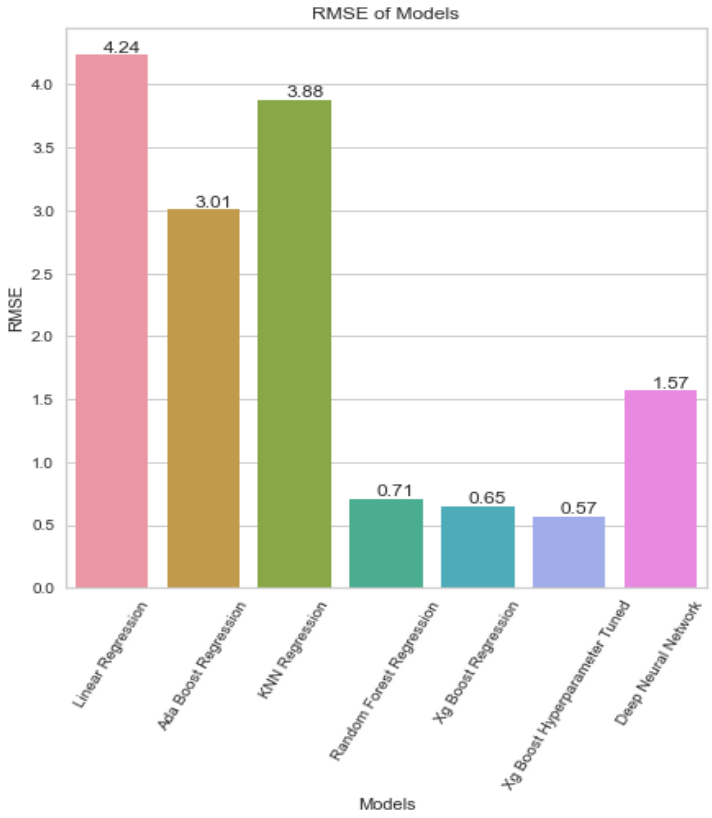


8. PERFORMANCE EVALUATION:

The mean Expected Lifespan is 14, hence the accuracy is calculated considering RMSE with respect to this.

After evaluating all the models the results came to be as follows:

MODEL	R2 SCORE	RMSE
Linear Regression	0.70	4.24
Ada Boost Regression	0.84	3.01
KNN Regression	0.52	3.88
Random Forest Regression	0.99	0.71
Xg Boost Regression	0.99	0.65
Xg Boost Hyperparameter Tuned	0.99	0.57
Deep Neural Network	0.91	1.57



9. CHALLENGES:

Some of the challenges faced during the process were:
Cleaning the categorical variables and removing the variants of the categories, deciding the amount of missing data to be dropped, considering the appropriate variables based on domain knowledge and obtained results and finally selecting the parameter values of deep neural network.

10. CONCLUSION:

After eveluating the performances of above mentioned models, it was found that machine learning algorithms like Random Forest Regressor and Xg Boost Regressor gave decent RMSE values. Also with the help of hyperparameter tuning the performnace was boosted.

The Deep Neural Networks gave excellent results. This also depends upon the parameters like the activation function, hidden layers, activation function, learning rate, optimizer, batch size and epoch.

At the end it all depnds on how well the data is structured to fit into particular model. Finally the choice of the model depneds upon how consistent and efficient the results suits the partcular application. So in order check the consistency of the results, stratified k-fold cross validation is carried out.

From the results we conclude that the Random forest, Xg Boost and Deep neural Network Models gave good performace and can be deployed for this particular application.

11. REFERENCES:

[1] Deep Learning wiith Tensorflow,

https://towardsdatascience.com/deep-learning- with-tensorflow-5d3a7a8c55cd