Rebecca Lersch
ISTA 421
December 16, 2025
Salena Ashton

Satellite Modeling

Satellites are a very important creation. They orbit the earth and provide many services to Earth below, without them life would be very different. The process to get a satellite into orbit is quite extensive. I found a dataset containing information on most active satellite's orbital information, and pre launch information. This dataset has 25 columns and over 1400 satellites. The columns are, Official Name of Satellite, Country/Organization of UN Registry, Operator/Owner, Country of Operator/Owner, Users, Purpose, Detailed Purpose, Class of Orbit, Type of Orbit, Longitude of Geosynchronous Orbit (Degrees), Perigee (Kilometers), Apogee (Kilometers), Eccentricity Inclination (Degrees), Period (Minutes), Launch Mass (Kilograms), Dry Mass (Kilograms), Power (Watts), Date of Launch, Expected Lifetime (Years), Contractor, Country of Contractor, Launch Site, Launch Vehicle, COSPAR Number, and NORAD Number. This dataset is important because having information about each and every satellite allows us to keep better track of where they are and any issues that may occur. This also allows for innovations to take place as there is historical information on satellites that can be used to investigate issues. To do this, specific information is needed about the satellite, so a problem needs to be identified. In this study, I am focusing on predicting the lifetime of a satellite, to see what information is the most important to make a satellite last longer. I will do this by using a linear regression model on numerical features and a logistic regression model on the categorical features. I hypothesize that in the linear model that the mass of the satellite will be the most significant feature, and the class of orbit to be the most significant feature in the logistic regression.

This problem of satellite lifetime is a very real issue. Once satellites reach the end of their lifetime, they essentially become "space trash." They are boosted into a different orbit and left there as waste. The issue in this lies in the ability of the trash to crash into one another. When orbital debris is created, it spreads entirely around the Earth in a matter of hours. The more debris spreading, the more debris that is created. This causes issues with active satellites that are carrying out missions, because now they are at risk. Orbital debris can also be seen in telescope observations which is a great issue, and can obscure scientific measurements. By aiming to understand what factors affect the lifetime of a satellite, we can work with the creators to better reduce our impact on the space environment. There is only so much room around the Earth, and if it becomes cluttered with waste, then we lose access to many services that have become essential to our daily lives, such as the internet, GPS, satellite TV, and many other things. There are many different types of satellites needed to provide those services. There are military satellites, government satellites, GPS, and communication satellites. Communication satellites are satellites that communicate information back to Earth such as internet or TV. Each of these types of satellites have a specific distance away from Earth that they need to be in order to do their best. There are four main orbital regimes, Low Earth Orbit (LEO), Mid/Medium Earth Orbit (MEO), Geostationary/ Geosynchronous Earth Orbit (GEO) and Cislunar Orbit. Along with different orbits, each satellite will have a relatively unique mass depending on the size and mission of the spacecraft. There are two forms of mass, total mass, and dry mass. Dry mass is all the solid mass, essentially everything except fuel or any other liquid. In order to get into orbit, a satellite must be on some sort of launch vehicle. That means a rocket needs to launch and bring

them into space. Some of the more popular launch vehicles in recent years are from SpaceX due to some relatively new technology that allows for the rockets to be reusable. This is a huge advancement in lowering the amount of orbital debris. It is important to understand specific characteristics of each satellite for characterization purposes, as well as to better understand how to make better satellites to reduce the amount of space waste.

There are two methods used in this study to predict the most significant feature in the lifetime of a satellite. The first being a linear model, where only the numerical information is used. This is likely to yield very actionable results since the information in this model is purely statistics of the satellites themselves. I hypothesize that launch mass will be the most significant feature, since heavier satellites need more fuel to operate and keep station keeping. With more fuel consumption, less time is available for the satellite to be active. In the logistic model, I will be focusing on categorical information such as class of orbit, owner, operator, country of origin, launch vehicle, launch site, etc. In this model lifetime of the satellite will be one-hot encoded to a predefined long lifetime threshold, with a binary indicator, for long lifetime or not. I hypothesize that the class of orbit will be the most significant feature. This is because the further that you go away from Earth, the longer it takes to go around, and more fuel consumption. While these are simply hypothetical, what it would mean if those are true, is that fuel consumption is a large factor in the lifetime of the satellite, and should be focused on more by the manufacturers of these satellites to increase the lifetime. Though, to keep in mind there is no information about solar panels on these satellites, which makes that outside the scope of this research, but something to keep in mind.

When using this data, from Kaggle, it needs to be useful for this study. Since there are going to be two models, two dataframes are needed to contain the data. Though, before that happens, we need to find the interesting information, and the unnecessary information. Unique data, such as name, and identification numbers are not going to be of use in any predictive modeling. That means we need to drop the columns with name, NORAD ID, and COSPAR number. NORAD ID and COSPAR are both systems for identification, one being incremental from the launch of the first satellite, and the other being a launch date system respectively. There are columns for the purpose of the satellite, which will be communication, military, etc, and also a detailed purpose. The detailed purpose is highly variable satellite to satellite and simply is not needed in this study, so detailed purpose will be dropped. Another feature that has been dropped is the Country/ Organization of UN registry. That is because most often it is repeat data found in other columns for each satellite. The next step is to filter out any rows where there are blanks in our Expected lifetime column. This needs to be done because that is the information we care most about. If there are blanks, then we cannot predict it. Then we split into our numerical dataframe and categorical dataframe. From there, we can drop any rows that have blanks. This leaves us with over 400 observations per dataframe. Some work had to be done in some of the columns using regular expressions because of things like ‘14yrs’ rather than ‘14’. Then we split each of the dataframes into test and train sets for their respective modeling. It is important to understand the distribution of years in the datasets. For the linear regression model a histogram showing the distribution is shown in Figure 1.

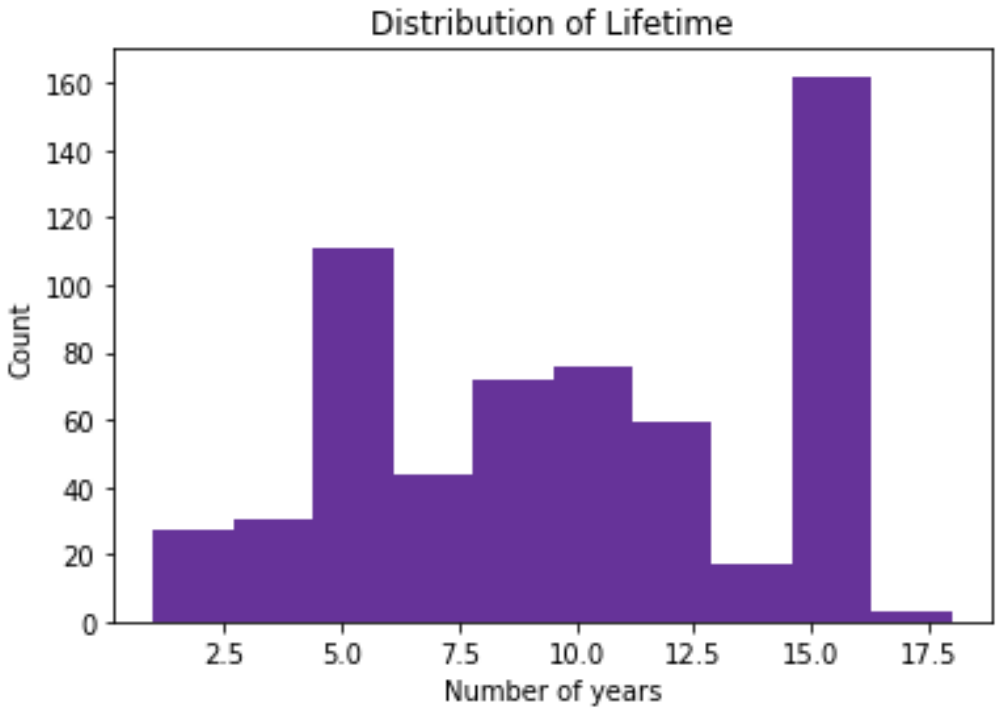

Figure 1.

We can see that there are some outliers, but mainly the lifetime is around 10 years. For the logistic regression model we can explore the same in Figure 2.

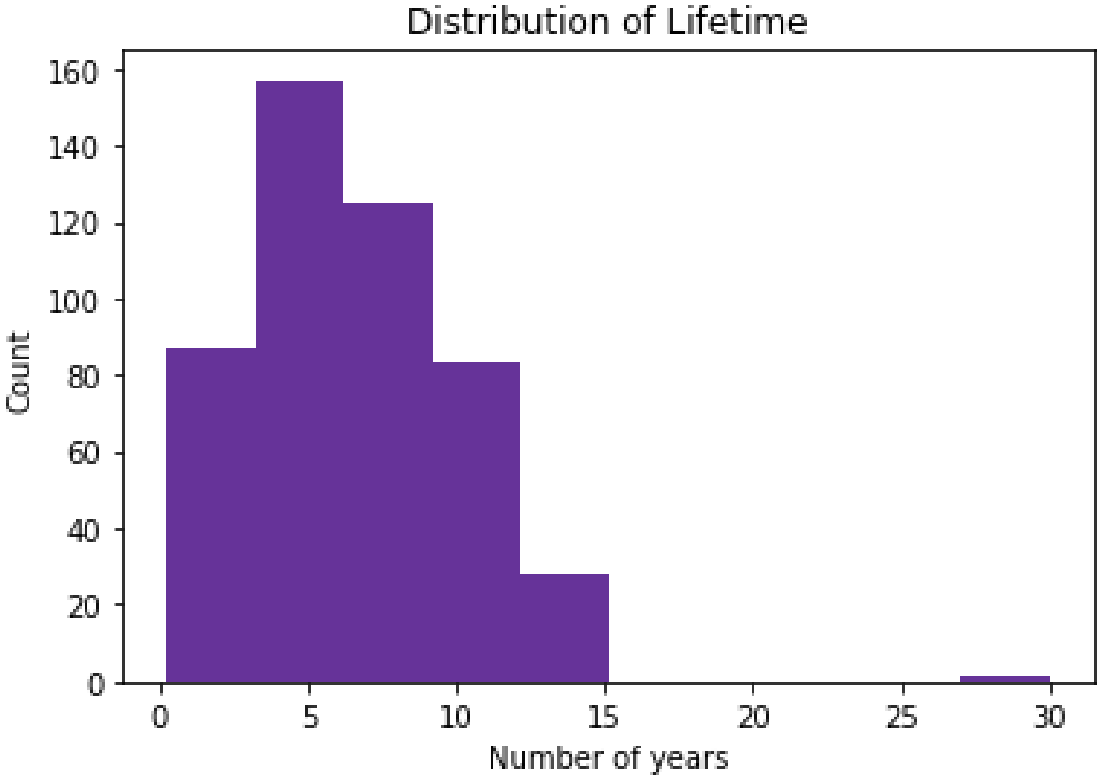

Figure 2.

We can see that there are more outliers, though generally about the same. Now, it is important to see how this conveys to our one-hot encoded column of long lifetime or not. In Figure 3 this is explored.

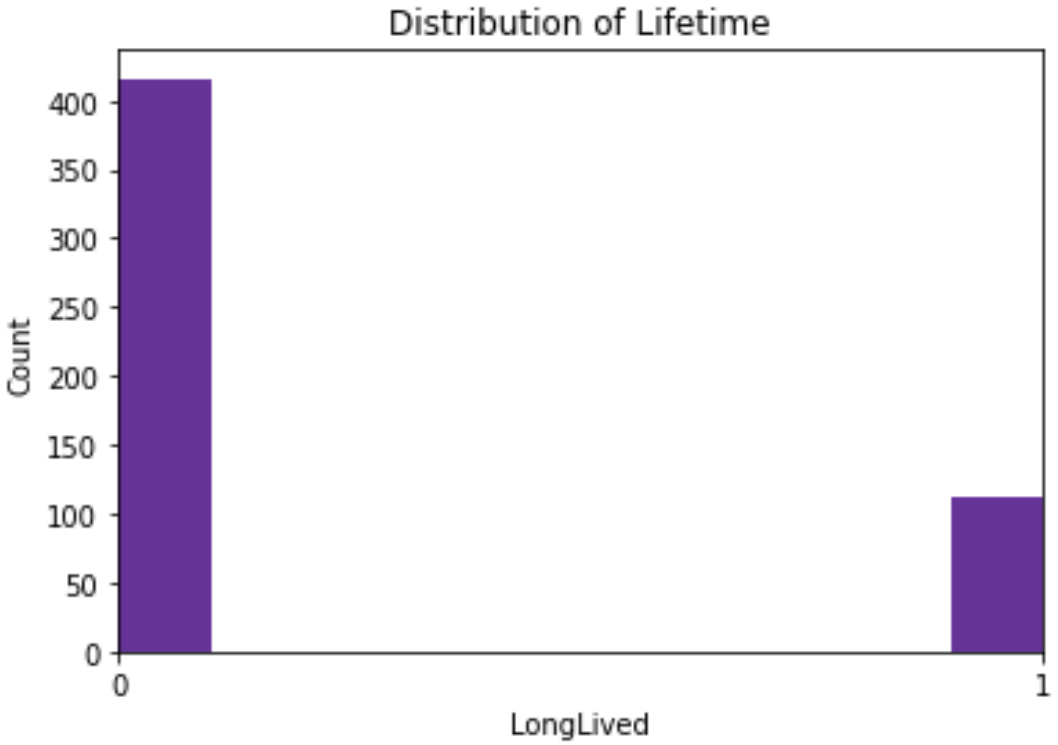

Figure 3.

We can see that the majority of satellites are not long lived, when we set our threshold to less than 10 years.

For the linear regression model, we have the data split into test and train. We can then work to fit a linear regression to this data. I have done this using the linear algebra library to assist in the matrix multiplication. This model was chosen, because it is the most easily interpretable by people for numerical data. It is also similar to logistic regression and allows for a more straightforward comparison for numerical and categorical data. Since the data has already been put into test and train, we can simply plug into the equation,

$$\hat{\beta} = ({X^T}_{train} X_{train})^{-1} ({X^T}_{train} Y_{train}) \ ,$$

And the equation for prediction being,

$$y_{test} = \hat{\beta}(X_{test}).$$

These can be put on a plot to visually access the model. In figure 4, we can see that.

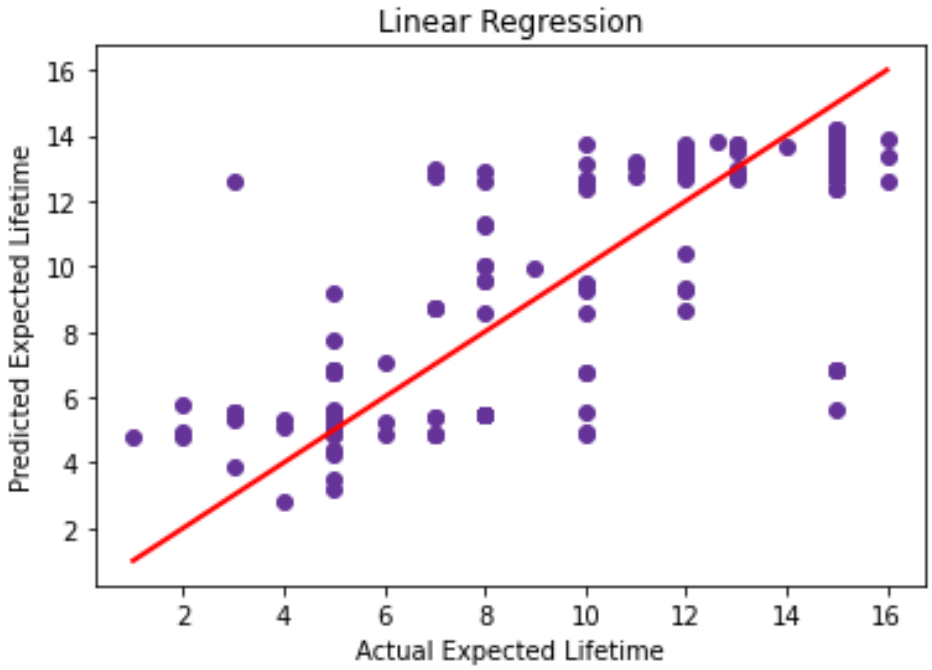

Figure 4.

In this figure, our visualization matches our regression line very closely, which helps indicate a good fitting model. The features chosen in this model have been picked from our dataset to specifically eliminate some collinearity. Specifically within Perigee and Apogee, since they are very related to each other, as they are the nearest point and furthest point from Earth in orbit. The dry mass was also removed because it is included within the Launch Mass. This model shows that the most significant feature from the data set is the inclination of the satellite in degrees. That is the tilt of the satellite's orbit in relation to the equatorial plane. We end up with an MSE: 8.5959, RMSE: 2.9319. This makes for a relatively decent model considering the number of features in the model.

These results from the linear regression model mean that the inclination of the satellite is the most important feature. The inclination being significant makes sense because of orbital dynamics becoming more complicated outside of Earth's equatorial plane. Overall, as the inclination increases, the lifetime of the satellite goes down according to the model.

Now that we also have the categorical features split into test and train, we can begin building the logistic model. I wrote a sigmoid function to turn my resulting logit into a probability, which is given by the equation,

$$\sigma = \frac{1}{1+e^{-z}},$$

Where Z is the logit. To find the logit, I wrote a logistic regression function that works by using gradient descent to iterate through and find better coefficients. Once applying these to my data, I found that the most significant feature was the owner/operator. This model ends up with Accuracy: 0.9524, Precision: 1.0000, Recall: 0.6875. This suggests that the model is fairly good at predicting if a satellite will be long lived or not. The results of the logistic regression are in figure 5.

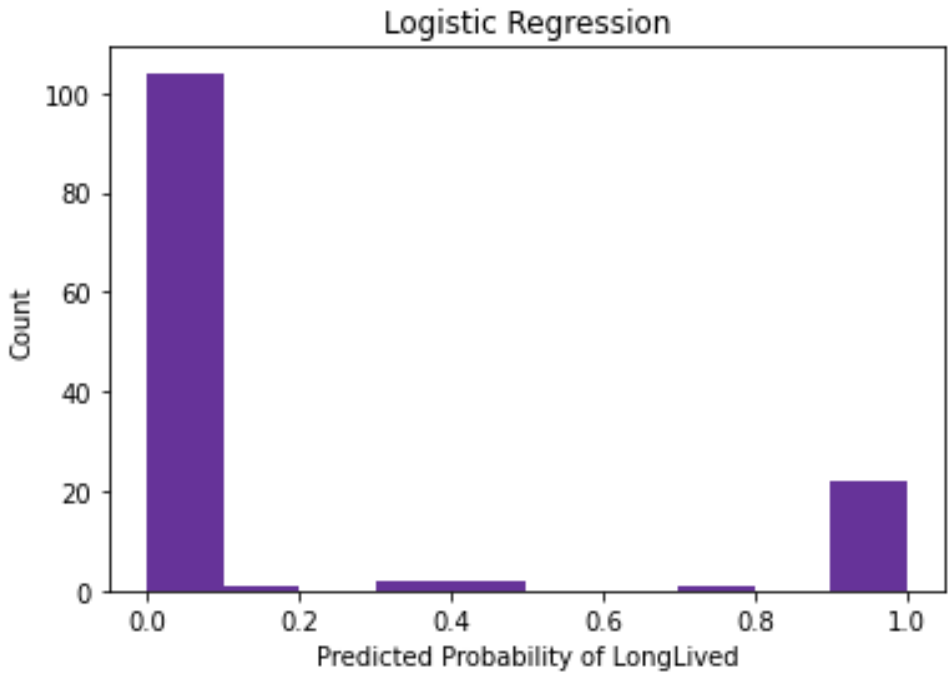

Figure 5.

By looking at the results of the logistic regression we see that owner/operator was the most significant feature. This could mean that specific countries specifically purchase satellites with specific features that are characteristic with longer or shorter lifetimes. Within this data and study will not know since there is no causal data, we can simply infer. Though, this outcome is slightly more difficult to interpret because of the lack of knowledge about each owner/ operator.

It is fairly apparent which model is better because of the interpretability. The logistic regression's most significant feature is fairly ambiguous in meaning and does not really allow for a solid, actionable outcome. Whereas the linear regression model, we can clearly understand how the feature might relate to longevity of the spacecraft. Inclination is an important feature because of the physics regarding how satellites move on an inclined orbit. This is actionable because manufacturers and operators can work together to find an ideal inclination to put a satellite that can balance lifetime and inclination.

Overall, the linear regression model is preferred because of the linear nature of the features and the interpretability of the outcomes. The features also are more actionable as previously stated. This is significant because using publicly available information, with this study I was able to determine what affects the lifetime of a satellite, and what to focus on when attempting to create a longer living satellite. A satellite with a longer lifetime is vital in sustainability that needs to take place in space. There are simply too many things in orbit around Earth and it is only going to get worse, especially with the increase in launches from companies like SpaceX or other places interested in creating mega constellations of satellites. Outside of sustainability, the amount of objects in orbit can cause issues for astronomers and investigating the cosmos, as the debris will get in the images and distort the data. I am surprised by the most significant features because they are different from what I expected in both models. The launch mass and inclination are not features that are related at all so this is very surprising to me. Same with class of orbit and contractor, they are not related at all.

In my future career or research, I can see how this will be very useful to know how to do. By being able to predict the outcome of an event using the data historically, I will be able to better give the full picture and story of the data. I am planning on entering the workforce sometime in the next 6 months, as a data scientist or data analyst. Machine learning is a very useful thing to be able to understand, even at an introductory level because of the widespread applications it has. This project in particular is useful, because I am interested in satellites and knowing how to take real world satellite data that I did not collect, and do some type of analysis on it to predict outcomes to make satellites better is interesting and fits into my current job.

Works Cited

"Linear Algebra." *Numpy Documentation*, https://numpy.org/devdocs/reference/routines.linalg.html.

"Mastering Logistic Regression: A Practical Guide with Python." *Medium*, https://python.plainenglish.io/mastering-logistic-regression-a-practical-guide-with-python-89b15374eb56.