Satellite Position and Satellite Velocity Prediction

Using CRISP-DM to develop a model to predict satellite position and velocity

Billy Jacob Franklin Jacob
Data Analytics
Dublin Business School
Dublin Ireland
10540532@mydbs.ie

INTRODUCTION

Satellite is a man-made object that has been intentionally placed into the Earth's orbit. There are 2,666 satellites currently orbiting Earth. The amount of satellites in Earth orbit is firmly growing and with the high amount of space debris, either crossing through or a resident in orbit, collision possibilities between two such objects can become hazardous. To avoid this scenario the prediction of the satellite position in space is necessary. As a part of this process, we will be clarifying the forecast of a Simplified General Perturbations-4 (SGP4) model. SGP4 can foretell a lot of effects but is applied to near-Earth objects with an orbital period of fewer than 225 minutes, while high flying orbit space objects have orbital periods up to 200 hours. For the actual position of the satellites, the position obtained using a more accurate simulator will be taken. Subsequently, the obtained models will be applied to real classified data and will help to predict the positions of these space objects. As a part of this report, the lessons learned throughout the progression of the CRISP-DM process will be provided in detail.

CRISP-DM

The cross-industry standard process for data mining is a standard process model that describes the common approaches used by the leading data mining institutions and industries al around the world. It is the most widely used analytics model.
CRISP-DM breaks the process of data mining into six major phases:

- Business Understanding
- Data Understanding
- Data Preparation
- Modelling
- Evaluation
- Deployment

BUSINESS UNDERSTANDING

In this phase, we need to understand the project objectives from a business perspective and then we should convert the knowledge gained into a data mining problem definition or a use case. Then, we have to develop the preliminary plan which is designed to achieve the objectives outlined earlier. As a part of this project, we need to clarify the prediction of a Simplified General Perturbations-4 (SGP4) model. Simplified perturbations models are a collection of five mathematical models (SGP, SGP4, SDP4, SGP8, and SDP8) used to calculate orbital state vectors of satellites and space debris relative to the Earth-centered inertial coordinate system. The simulated coordinates and velocities of the satellites are provided, and our objective is to predict the actual coordinates and the velocities of the satellites in the Earth’s orbit.

Project objectives: Design a model which can predict the actual position and velocity of the satellite in the Earth's orbit. The SMAPE score of the model should be less than 0.05. the model will be accepted if the SMAPE score is less than 0.05

Data mining problem definition: Design a model which can predict the actual position and velocity of the satellite in the Earth's orbit corresponding to the simulated positions and velocities of the satellites. The model with the least errors will be selected as the Final model.

Preliminary plan:
1) split the dataset into train and test data.

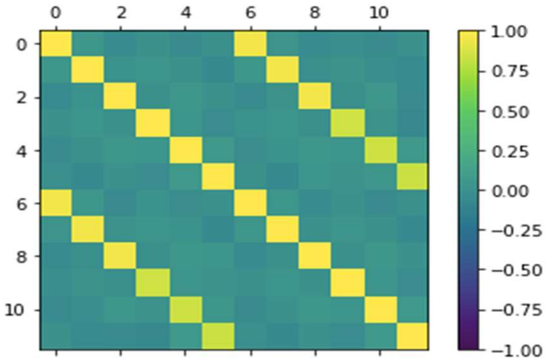
2) Find the relationship between the dependent variables and the independent variables using the train data.
3) Find the model which can predict the dependent variables with the least error using the test data.
4) Test the model in a new dataset that was not trained earlier.
5) Capture the results.

DATA UNDERSTANDING

In this phase, there are two steps that are collecting the data and increasing the familiarity with the data to identify data quality problems if present.
Data Description:
In the Dataset we have 649912 records and 15 attributes
The following are the attributes
id: This is the row id
epoch: The time at which the position of the satellites was recorded.
sat_id: The ID to differentiate each satellite. This dataset has the details of 600 satellites
x - actual x coordinate (to be predicted)
y - actual y coordinate (to be predicted)
z - actual z coordinate (to be predicted)
Vx - actual Velocity of the satellite in the x-direction (to be predicted)
Vy - actual Velocity of the satellite in the y-direction (to be predicted)
Vz - actual Velocity of the satellite in the z-direction (to be predicted)
x_sim - simulated x coordinate (independent variable)
y_sim - simulated y coordinate (independent variable)
z_sim - simulated z coordinate (independent variable)
Vx_sim - simulated Velocity of the satellite in the x-direction (independent variable)
Vy_sim - simulated Velocity of the satellite in the x-direction (independent variable)
Vz_sim - simulated Velocity of the satellite in the x-direction (independent variable)
There are six dependent variables and six independent variables in this dataset.

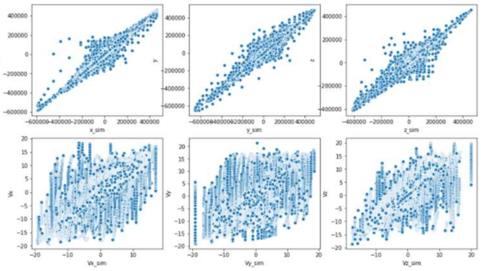
Data Exploration:
All the 600 satellites have 1-month data where the positions are provided for around 30 different times for example from 00:00:00.000 to 23:21:30.015 in a single day. All the data provided in the dataset includes the data from the year 2014 for the month of January.

Correlation

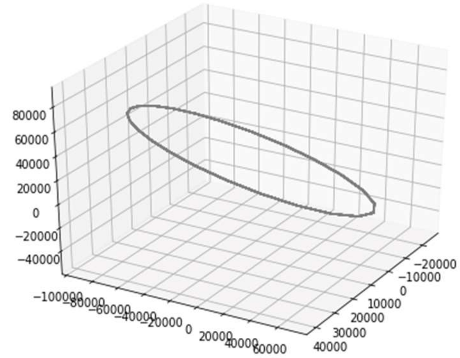


We found that there is no correlation between the feature and there is a negative or positive correlation only between the features and the dependent variables.

Scatter Plot:



3D plot of the coordinates



This 3D plot shows us that the points are in an elliptical shape along the orbit of the earth.

Verifying Data Quality:
The data is complete and there are no missing values in the dataset.
All the values are numerical, and all the records are unique as they denote the positions of the satellites in outer space.

DATA PREPARATION

This phase will contain all the activities to construct the dataset that will be fed into the data modeling tools.

Select Data:

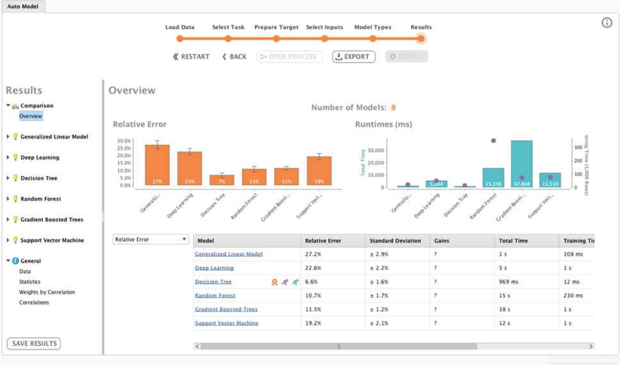
The sat_id and the epoch attributes will not be considered for the analysis of the data as they are completely irrelevant to our data mining goals. So, the final dataset will have 13 attributes id attribute, 6 dependent variables, and 6 independent variables. We are using the satellite id 0 for our modeling the 1st 30 days data is taken as the training dataset and the 31st day's data will be predicted using the 6 independent variables.

Clean data and the construct data steps are irrelevant in this dataset because there are no data quality issues and there is no requirement of new derived attributes or records. There is no change required to fit the needs of the modeling tools. Data Formatting is also not required in this dataset as there are no illegal terms, illegal characters, etc.

MODELING

Application of the auto model:

In this phase various modeling techniques are selected and applied.
We found that the decision tree was giving the results with the least error but when we tried the decision tree algorithm with new records the model was not able to predict.



Model	Relative Error	Standard Deviation	Gains	Total Time	Training Ti
Generalized Linear Model	27.2%	± 2.9%	?	1 s	108 ms
Deep Learning	22.6%	± 2.2%	?	5 s	1 s
Decision Tree	6.6%	± 1.6%	?	969 ms	12 ms
Random Forest	10.7%	± 1.7%	?	15 s	230 ms
Gradient Boosted Trees	11.5%	± 1.2%	?	38 s	1 s
Support Vector Machine	19.2%	± 2.1%	?	12 s	1 s

Selection of Modeling Technique

We selected the Neural Network for predicting the dependent variables. Neural networks are a series of algorithms that mimic the operations of a human brain to recognize relationships between vast amounts of data. Neural networks will be used to find the coefficients for each of the attribute using the activation function and the Back-propagation method. The neural network contains 2 hidden layers and 4 nodes each. We have captured some assumptions during this selection. We are assuming that there is relation between the dependent variables and the independent variables and applying the optimum coefficients to the independent variables will give us the dependent variables.

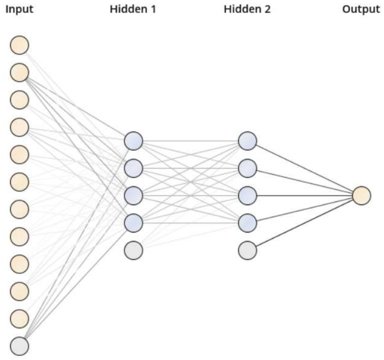


Figure 1: Figure of the Neural network generated after running the model in Rapid Miner.

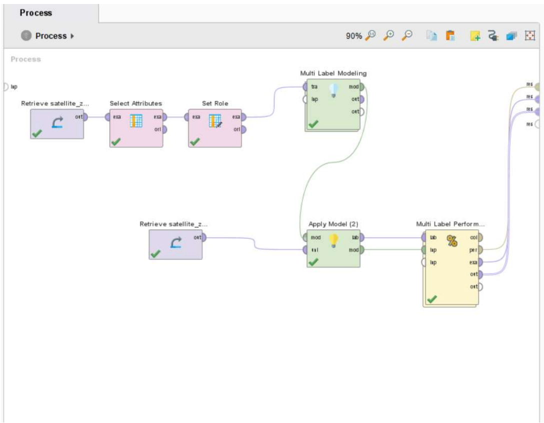


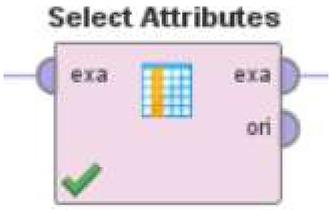
Figure 2: Figure of the Design in the Rapid Miner studio.

Sampling:

There was no specific sampling technique involved in this modeling step because all the values are unique and are coordinates of the satellite so there is no involvement of a bias in the supervised learning process.

Select Attributes:

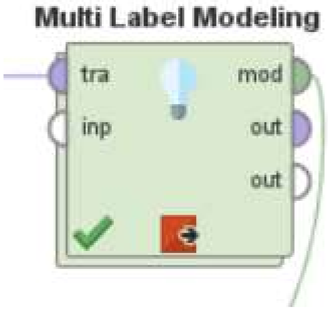
In the select attributes step all the attributes are selected except the SAT_ID and EPOCH attribute in the dataset.



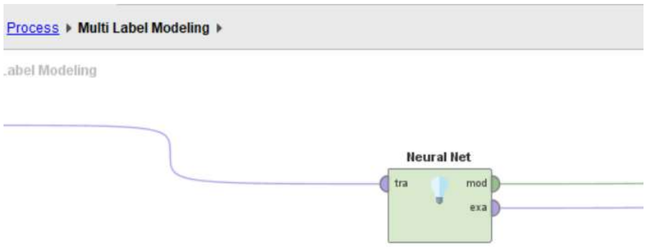
Set Role:

The ID attribute is set to id in the set role step.

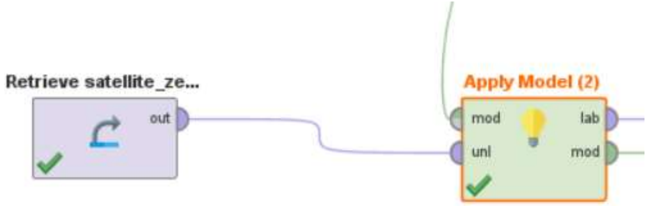
Multi Label Modelling:



The multi label modeling is used to set the label role to six attributes in a single run. If not, we have to create six different models for the six different attributes. Inside the multi label modelling operator we have the Neural Network

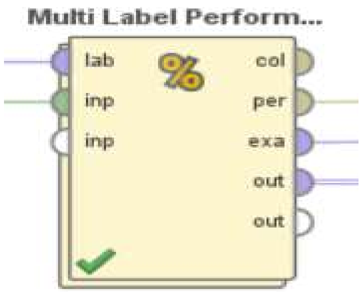


Apply Model to the test Data:

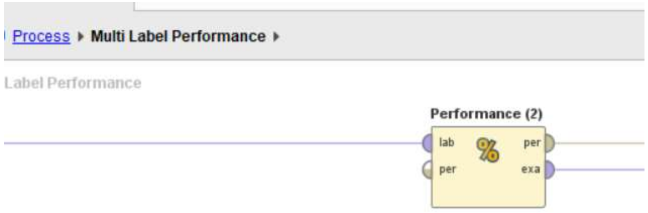


Here we apply the selected model to the test data which has the data for the 31st day of the month.

Multi Label Performance:



Here we have the multi label performance operator to find the performance of the model by taking into account all the 6 predicted attributes at the same time.



Inside the Multi label performance operator we have the Performance (Regression) operator which will be used to find the performance of regression models. As our use case is something related to the regression analysis, we will be using Performance (Regression) operator.

EVALUATION

In this phase the model is thoroughly evaluated before proceeding to the final deployment. The model was tested using three different algorithms and we found that the neural network algorithm was able to produce the results with the least errors.

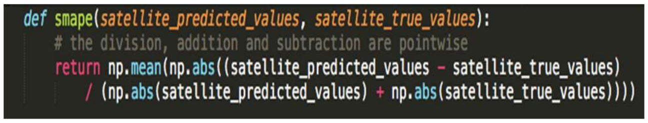
Evaluate results

	RMSE	Absolute error	Normalized absolute error
Linear Regression	556.870	439.826	0.075
Polynomial Regression	997689103149	806342646655.4	460718191045
Neural Networks	180.892	150.098	0.019

From the above comparison we can come to conclusion that Neural Networks are the best algorithms for this use case

SMAPE Score

The SMAPE (Symmetric Mean Absolute Percentage Error) is a variation on the MAPE that is calculated using the average of the absolute value of the actual and the absolute value of the forecast in the denominator. This statistic is preferred to the MAPE by some and is usually used as an accuracy measure in several forecasting competitions. We are also using this score to evaluate our model. We used the flowing piece of python code to find the SMAPE score for the predicted results.



```
def smape(satellite_predicted_values, satellite_true_values):
    # the division, addition and subtraction are pointwise
    return np.mean(np.abs((satellite_predicted_values - satellite_true_values)
        / (np.abs(satellite_predicted_values) + np.abs(satellite_true_values))))
```

The following SMAPE scores were obtained for the dependent variables.

x – 0.0028
y – 0.0029
z – 0.0025
Vx – 0.0014
Vy – 0.0010
Vz - 0.0010

At the end of the evaluation process we are supposed to determine the next steps here the project leader must decide whether to finish this project and move on to deployment or whether to initiate further iterations or to set up new project cycle. As we reached the required goal and objective, we decide to move on to the deployment process.

DEPLOYMENTS

In this step the knowledge gained from all the above steps should be organized and presented in a way that it will be useful and purposeful for the end user or the customer. Usually the models are deployed as live models within an organization for the prediction or decision making purpose.

Application of our model:

The model can be deployed in the internal server of any space research organization where the simulation of the satellite positions and velocities are modeled using the SGP-4. When the simulated coordinates and velocities are provided as input to the model the predicted velocities and the coordinates will be provided as output from the model.

Lessons Learned from this Activity:

- If we have the proper business understanding Data mining models could be applied in a vast variety of fields.
- It is very important to determine the goals and objectives at the beginning on the project so that those factors can be used during the evaluation phase.
- Data quality won't be the same all the time. The dataset which we selected for this project had very less data quality issues.
- Auto model can be deceptive at times as the model which have the best performance in the train data does not mean that they would have good performance when implemented on test data.
- Learned about the multi label modelling operator where more than one labels can be trained at the same time.
- Learned about the multi label performance operator where performance of more than one labels can be calculated.
- Learned about SMAPE score which are most widely used in forecasting competitions and hackathons.
- Learned a lot about Neural networks and how having different nodes affect the prediction quality of the model.
- Learned about the necessity of cross validation.