DATA SCIENCE AND ANALYTICS DISSERTATION

# Classification of bow shock and Magnetopause event positions by using the Magnetic Field properties of Saturn

A thesis submitted in partial fulfillment of the requirement for the degree of M.Sc. in Data Science and Analytics.

Author: Mathews Philip Venattu (20250487)

Supervisor: Dr. Katarina Domijan

Department of Mathematics and Statistics
University: National University of Ireland, Maynooth.

Date Submitted: 10th July, 2021

# Abstract

Cassini-Huygens Space Research mission was a joint collaboration of NASA, ESA and Italian Space agency to study about the planet Saturn and its system that includes its Rings and Natural Satellites. During its life span of about 20 years the spacecraft orbited the planet Saturn for 13 years and also frequently changed its shape and tilt. Because of this frequent changes in the orbital trajectory of spacecraft, it passed multiple times through the boundaries of Bowshock and Magnetopause at different latitude, longitude and phases of Solar Cycle. In this Project, we use the Magnetic Field data collected during the year 2005 from MAG (Magnetometer) instrument attached to the spacecraft to develop a classification model for detecting the Magnetopause and bow shock boundaries at saturn . MAG (Magnetometer)in Cassini recorded the strength and direction of the Magnetic field around the spacecraft while it was orbiting the planet Saturn.

# Contents

# List of tables

# List of figures

# 1 Introduction

All planets in our solar system have magnetic field like earth. Some of the planets like Uranus, Saturn, Jupiter, and Neptune has large magnetic field than earth. Magnetosphere of an astronomical object is the area surrounding that object where its Magnetic field is present. Like all other planets Saturn also has the similar magnetosphere structure - it has a Bow shock, Magnetosheath, Magnetopause and Magneto tail. Among this structure bow shock is the point at which the Magnetosphere of the Saturn interacts with the solar wind that in turn suddenly reduces its speed and pressure. Magnetopause is the boundary between Saturn's Magnetic field and Solar wind. The Magnetosheath exists between the bow shock and the Magnetopause, an area of shocked solar wind that is significantly influenced by the changes that occur within the bow shock and whose features can impact the interaction across the Magnetopause. The dynamic pressure of Solar wind usually determines boundary and position of Magnetopause and Bow shock [2]. The Cassini Huygens Mission is a joint NASA/ESA/ASI project to make a detailed survey of the ringed planet Saturn and its natural satellites. Cassini spacecraft recorded magnetic field and plasma condition of the environment during its insertion to Saturn's orbit by using the Cassini Magnetometer (MAG), Plasma Spectrometer (CAPS). The magnetic field strength pattern clearly shows some "overshoot" and "foot" when the spacecraft passed through the boundaries of Magnetopause, Bow shock and Magnetosheath [3]. The Magnetometer (MAG) which continuously acquired magnetic field data that is associated with the Plasma Environment and internal source of Saturn are essential to study about the interaction between solar wind and Magnetosphere of Saturn.

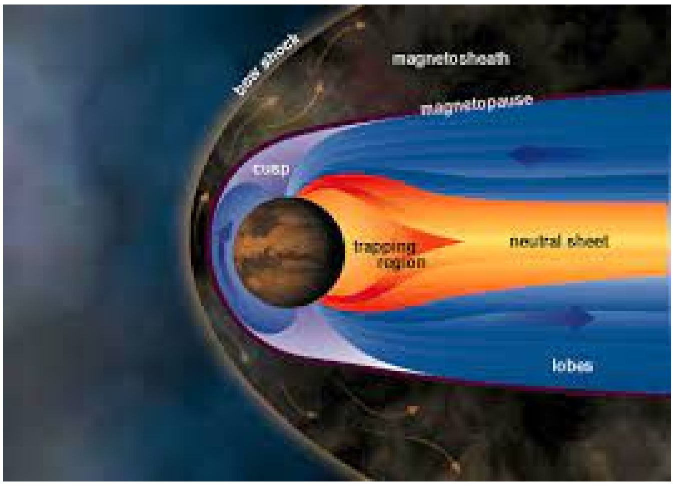

Figure 1: Diagram of Planetary Magnetosphere

####Explain the figure here

Magnetosphere of Saturn act as an obstacle to the Solar wind far away from the planet and the Magnetopause of the planet Saturn lies in 20 Rs (Radius of Saturn=60330 km). So, the Solar wind which interact with the magnetosphere is decelerated by the bow shock and the physical properties of the plasma (Sergis (2013)). The magneto disk pressure, which inflates the equatorial magnetosphere considerably more than the high-latitude magnetosphere, affects the geometry of the Magnetopause boundary itself, resulting in clear polar flattening [5]. In the case of Earth, basic pressure balance is due to the interaction between upstream solar wind flow against magnetic pressure from the magnetosphere and this helps to draw the approximate location of Magnetopause boundary but when it comes to Saturn we must consider the influence of the natural satellite Enceladus, which serves as a huge internal plasma source. The pressure related with the super thermal component of this internally produced plasma serves to inflate the magnetosphere considerably beyond what a basic magnetic pressure calculation would predict [6]. So even in a steady solar wind conditions, Due to the internal plasma mentioned above the Magnetopause boundary of Saturn may move significantly.

Lots of studies were done based on the data acquired from the Cassini Spacecraft instruments. (Sergis (2013)) they chose intervals between 2004 and 2011 when the spacecraft was travelling through the magnetic sheath and used the data recorded to explore different properties like plasma, energetic particle, magnetic field density, temperature etc. They concentrated more on the presence of W+s ions (water group) and explained the ejection of energetic ions as a function of pitch angle and energy which shows the often flow of heavy energetic ions from bow shock. Analysis of CAPS (Cassini Plasma Spectrometer) by Burkholder et. al [8] shown the significant ion flow differences prenoon and post noon, and apart from the magnetic field data was used to illustrate the rotation of IMF (Inter Planetary Magnetic Field) vector.

In this project we are considering the bow shock and Magnetopause crossing of the spacecraft during the year 2005. This report explains about the entire project with different sections like Background, Dataset explanation, Data Manipulation, Data Visualization, Models trained to classify the type of crossings, R code used, summary of the results and conclusion. Background section explains about the different terms used and current approaches done by others for classification. In Dataset explanation section it explains about each variable and the different datasets that were used for this project. Data Manipulation section gives a detailed explanation about the transformations and imputations done on the dataset. New dataset made after data manipulation is explained visually on the data visualization section. There are different models tried to achieve better results each model that are used for this project is explained on this part of the report. Remaining portions explains about the code and its results.

# 2 Background

In the early days, scientists have very less information regarding the Planet Saturn and its magnetosphere because as we all know that the magnetic fields are invisible, and it needs to be studied from inside. Cassini Huygens mission was a great opportunity for the science world to explore the planet and its behavior. This mission helped to study the magnetic field and the flow of different gases under the influence of the magnetic field which affects the auroras of Saturn. This mission given some powerful insights about the atmosphere and the surrounding of Planet Saturn. By comparing Saturn with the similar exoplanets will give information regarding the evolution of the Solar System. Different studies were done based on the data gathered from the Cassini spacecraft. Based on this data [9] demonstrated that polar flattening of the Magnetopause causes shorter streamline pathways over the poles, resulting in a higher-pressure gradient, which twists the field. This in turn leads to different conditions at the Magnetopause when compared to those predicated based on axisymmetric assumptions. From 2004 day 299 through 2012 day 151, a substantial data was compiled by [10] of Magnetosheath measurements was collected using data from CAPS, MAG, and MIMI. This data collection enables researchers to investigate things like local temporal dependence of Magnetosheath parameters. They also demonstrated a new method for estimating upstream solar wind speed using the same Magnetosheath parameters. [11] used the MAG data for research which provides a broad picture of low-frequency waves in Saturn's magnetosphere, which has crucial consequences for how magnetospheric energy leaks.

Both the Bow shock and Magnetopause models can be used as a significant tool which gives insights about the solar upstream conditions and its dynamic pressure at which they are associated. All the data associated with Cassini uses KSM coordinate system and this system is Saturn centred where the x -axis is towards the sun [13]. Orbital tour of Cassini around Saturn which started in the month of July 2004 during that time the spacecraft crossed 100 Bow Shock boundaries. A study done by [12] On 11th and 12th of April 2005, Cassini magnetometer readings were made during a typical sequence of Cassini bow shock crossings. The spacecraft began and finished the period downstream of the shock in the Magnetosheath solar wind, with two trips into the upstream solar wind, each separated by two shock crossings. The presence of obvious shock ramps and a constant upstream field indicates that these are quasi-perpendicular crossings. During this time magnetic field strength values recorded by the magnetometer were so high. In this report I analysed the magnetometer and position data of Cassini spacecraft during the year 2005 to classify the Magnetopause and Bow Shock events.

# 3 Datasets

Cassini Spacecraft orbited around the Saturn for about 9 years.During this period the spacecraft transmitted valuable information regarding Saturn like the magnetic field strength, position at which it was measured to earth. Magnetometer and CAPS are the main instruments that were used for measuring the magnetic field strength and Kinetic Energy of particles at each point. For this project, I am only considering the data that was recorded during the year 2005 by the spacecraft. Mainly two datasets were used in this project to make a final combined useful dataset, first dataset contains a list of Bow Shock and Magnetopause event crossings that occurred during the year 2005 (Jackman et. al,2019). The second dataset contains the information regarding the position of spacecraft and the vector data of Magnetic field strength.

## 3.1 Raw Datasets

### 3.1.1 Dataset 1: Magnetopause And Bowshock Crossing List

This dataset only contains the data of the year 2005 and that was originally developed by compiling two datasets that are posted in the MAPSView webpage (http://mapskp.cesr.fr/BSMP/index.php) which contains the Bow Shock and Magnetopause event crossings between 2004 day 179 and 2007 day 349 (H.J. McAndrews, S.J. Kanani, A. Masters, and J.C. Cutler) through visual identification of CAPS and MAG data. The second list of data has the Magnetopause crossings during the year 2004 to October 2010 and May 2012 to February 2013 [6].

This dataset contains seven variables: `year_cross`, `doy_cross`, `doyfrac_cross`, `hour_cross`, `minute_cross`, `type_cross`, `dirn_cross`, `xcrosslist`, `ycrosslist`, `zcrosslist`

Variable Description

- `year_cross`: It contains a numeric value of the year in which spacecraft crossed the event.
- `doy_cross`: It contains a numeric value of the day on which spacecraft crossed the event.
- `hour_cross`: It contains a numeric value of the hour at which spacecraft crossed the event.
- ‘minute_cross ‘: It contains a numeric value of the minute at which spacecraft crossed the event.
- `doyfrac_cross`: `doy_cross` + (`hour_cross`*60+minute_cross)/(24*60)
- `type_cross`: This is a categorical variable contains information about what type of event did the spacecraft crossed.
  - **MP**: Magnetopause
  - **BS**: Bow Shock
  - **DG**: Data gap
  - **SC**: SCAS interval which are unreliable data
- `dirn_cross`: This is also a Categorical variable that contains information regarding in which direction did the spacecraft moved.The direction categories in this variable are:
  - **in**: Inbound means the spacecraft is moving towards the planet.
  - **out**: Outbound means the spacecraft is moving away from the plant.
  - **S_SW**: Starts with the solar wind is the region at which spacecraft recorded values at the start of solar wind.
  - **S_SH**: Starts with Magnetosheath is the region at which spacecraft recorded values at the start of Magnetosheath.
  - **S_SP**: Starts with Magnetosphere is the region at which spacecraft recorded values at the start of magnetosphere.

- **E__SW**: Ends with the solar wind is the region at which spacecraft recorded values at the end of solar wind.
- **E__SH**: Ends with Magnetosheath is the region at which spacecraft recorded values at end start of Magnetosheath.
- **E__SP**: Ends with Magnetosphere is the region at which spacecraft recorded values at the end of magnetosphere.

For an inbound the first event that will occur is a Bow Shock and later followed by Magnetopause. But in the case of an outbound direction the first event that occur will be a Magnetopause and later followed by a Bow Shock. Region of sampling at the start of any data gap will have a S__SW, S__SH and S__SP direction type and Region of sampling at the end of data gap E__SW, E__SH and E__SP. Dimension of this dataset is (480,10).

### 3.1.2 Dataset 2: Magnetometer Dataset

In this dataset it contains the magnetometer data of Cassini Spacecraft during the year 2005. Time difference between each Data points is one minute which means each data point represents the data of a particular minute. The data in this dataset are provided in the KSM (Kronocentric Solar Magnetospheric) Coordinate system which is a kind of Saturn centred Coordinate system where direction of X is from Saturn to the Sun and X-Z plane of the Coordinate system contains the Saturn centred axis of Magnetic Dipole 'M'.

Some of the relevant variables in the dataset are:

- `X_KSM.km.`: This is the X coordinate point value of the spacecraft in KSM Coordinate System.
- `Y_KSM.km.`: This is the Y coordinate point value of the spacecraft in KSM Coordinate System.
- `Z_KSM.km.`: This is the Z coordinate point value of the spacecraft in KSM Coordinate System.
- `Timestamp.UTC.` : It is the timestamp at which data point was recorded by the Magnetometer.
- `DOY.UTC.` : It tells about the day at which the datapoint was recorded in the year 2005.
- `BX_KSM.nT.`: It is the x component of magnetic field strength in Amperes/meter.
- `BY_KSM.nT.`: It is the y component of magnetic field strength in Amperes/meter.
- `BZ_KSM.nT.`: It is the z component of magnetic field strength in Amperes/meter.
- `BTotal.nT.` : It is the resultant vector of Bx, By and Bz

$$B_{Tot} = \sqrt{Bx^2 + By^2 + Bz^2} \tag{1}$$

This dataset contains 494683 rows and 12 columns.

## 3.2 Derived Datasets

This section contains the information regarding all the datasets that were derived from the Raw Datasets.

### 3.2.1 Dataset 3: Combined Data Of Dataset 1 And Dataset 2

This is a newly created dataset by merging dataset 1 and dataset 2 So, that we can understand the Magnetic field properties during the events like Magnetopause and Bow Shock. For merging the two datasets I used date and time as the key. To format the date in Dataset 1 I used the `doy_cross` variable in each row adding to "2004-12-31" date. By doing so it will generate a date with respect to the reference date. In dataset 2 the variable `Timestamp.UTC.` is in string format inorder to convert it into a data format I used `as.POSIXct()` function with `format="%d/%m/%Y %H:%M"`. Later I have converted the `Timestamp.UTC.` variable into a new format and which is then stored in the `date` variable. Two new variables are also created in the Dataset 2 known as `hour_cross` and `minute_cross`. `left_join()` function was used for merging the two datasets by using the variables `date`,`hour_cross` and `minute_cross` which is common on dataset 1 and dataset 2. In the newly created dataset it contains all the variables of dataset 1 and dataset 2. The dimesnsion of the newly created dataset is 494683 rows and 19 columns. Some of the variables are removed from the dataset because we know that dataset 1 has very less number of datapoints when compared to dataset 2 So, it is better remove the variables like `xcrosslist`, `ycrosslist`, `zcrosslist`, `year_cross`, `doy_cross`,`SCET.s.`, `doyfrac_cross`, `hour_cross` and `minute_cross` from the merged dataset. Data manipulations and Visualizations were done on this newly created dataset.

After removing some of the variables, now the modified dataset hass 494683 rows and 16 columns.

### 3.2.2 Dataset 4: Average and Standard Deviation Dataset

This dataset was created after the exploratory data analysis done on `Dataset 3`. From the results of those analysis Standard deviation and Average value of the Total Magnetic field 15 minutes before and 15 minutes after of an event occurred data point are showing some pattern.

**Some of the Variables are**

- `Avg_Lag_Bx.`: This is the Average of all the `Bx` values that were recorded 15 minutes before each datapoint.
- `Avg_Lag_By`: This is the Average of all the `By` values that were recorded 15 minutes before each datapoint.
- `Avg_Lag_Bz`: This is the Average of all the `Bz` values that were recorded 15 minutes before each datapoint.
- `Avg_Lag_BTot`: This is the Average of all the `BTot` values that were recorded 15 minutes before each datapoint .
- `SD_Lag_Bx.` : This is the Standard Deviation of all the `Bx` values that were recorded 15 minutes before each datapoint.
- `SD_Lag_By` :This is the Standard Deviation of all the `By` values that were recorded 15 minutes before each datapoint.
- `SD_Lag_Bz`: This is the Standard Deviation of all the `Bz` values that were recorded 15 minutes before each datapoint.
- `SD_Lag_BTot.`: This is the Standard Deviation of all the `BTot` values that were recorded 15 minutes before each datapoint.
- `Avg_Lead_Bx.`: This is the Average of all the `Bx` values that were recorded 15 minutes after each datapoint.
- `Avg_Lead_By`: This is the Average of all the `By` values that were recorded 15 minutes after each datapoint.

- `Avg_Lead_Bz`: This is the Average of all the `Bz` values that were recorded 15 minutes after each datapoint.
- `Avg_Lead_BTot`: This is the Average of all the `BTot` values that were recorded 15 minutes after each datapoint.
- `SD_Lead_Bx.` : This is the Standard Deviation of all the `Bx` values that were recorded 15 minutes after each datapoint.
- `SD_Lead_By` :This is the Standard Deviation of all the `By` values that were recorded 15 minutes after each datapoint.
- `SD_Lead_Bz`: This is the Standard Deviation of all the `Bz` values that were recorded 15 minutes after each datapoint.
- `SD_Lead_BTot.`: This is the Standard Deviation of all the `BTot` values that were recorded 15 minutes after each datapoint.

# 4 Data Manipulation Section

In this section will explain about the data manipulation that was done on the merged dataset (dataset 3). There were lots of NA values in different predictors, so it is important to impute these values before using it for training the models.

```
## Warning: package 'caret' was built under R version 4.0.5
```

## 4.1 Removing Time Dependency

Since all the datapoints were recorded by the spacecraft using the instruments over time so, there can be a time dependency. To remove the time dependency, I made the dataset wider which means a thirty-minute window was used for each data point and stored the magnetic field strength values and position of the spacecraft at each minute as a column for each row. Now for each data point there are 219 columns. I have labelled each column in the format (predictor_name{minute_index}) For example, the BX_KSM16 represents the BX_KSM value after one minute of the selected datapoint. Since all time-dependency variables need to be removed So, in this dataset the variable Timestamp.UTC. ‘ was removed for this purpose.

## 4.2 Data Imputation

In this dataset there are many NA values in different predictors like `type_cross` and `dirn_cross`. Since `type _cross` and `dirn_cross` are both categorical variables So, it is critical to impute the NA values with relevant short terms. `type_cross` variables represents the type of event at which the spacecraft crossed. Currently `type_cross` variable has values `MP`, `BS`, `DG` and `SCAS` which represents Magnetopause, Bow shock, Data gap and Unreliable data. All the other data than the above-mentioned categories in the newly created dataset can be categorized as `NE` which means No Events Occurred.

For `dirn_cross` variable, which represents the direction at which spacecraft is moving. This variable has the categories E_SH, E_SP, E_SW, I, O, S_SH S_SP and S_SW. So, I have imputed all the datapoints which has NA values in `dirn_cross` as `UD` (Unknown Direction) which means the direction of the spacecraft when that datapoint was recorded is Unknown.

## 5 Exploratory Data Analysis

The dataset contains the records of more than 490000 magnetometer readings with labels of type of crossing, Magnetic filed strength values of fifteen minutes before and after of a datapoint and position of the spacecraft at which the data was recorded. The orbits of the spacecraft covered almost all local hours and gave sufficient dayside coverage. Before getting into further analysis its important to understand whether the data is imbalanced or not.The dataset contains enough datapoints for training different models. Understanding the each variables on data is required before training the Models for classifying the events. Eventhough dataset contains

To understand the inbound and outbound boundary crossings. From [1] It is stated that from 72th day to 74th day of the year 2005 Cassini Spacecraft was in Outbound on the dawn flank. Figure shown below is the Line graph of Total Magnetic Field recorded during the period of 72 to 74th day of the year 2005. In the Figure below the blue dotted line represents the point at which spacecraft crossed Bow Shock boundary and the red dotted line represents the Magnetopause.

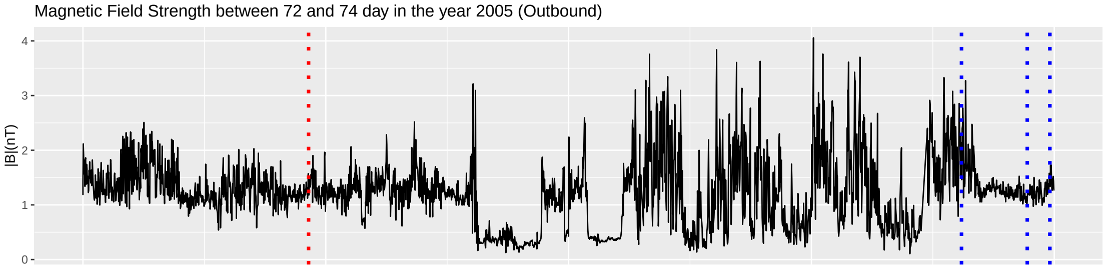

Figure 2: Line Graph of Total Magnetic Field Strength between 72 and 74th day of the year 2005

From the above figure we can clearly see that when the spacecraft approached the Bow Shock boundary there is a large value fluctuation of total magnetic field but when it comes to Magnetopause event there are only small value fluctuations. Since this data was recorded during the Outbound, the first boundary that was crossed by the spacecraft was Magnetopause and later followed by the Bowshock Boundaries.

During the Inbound prenoon of Cassini Spacecraft which is between 136th and 138th day of the year 2005, the spacecraft observed some clean bow shock crossings on 136th day and followed by the Magnetopause Crossings on 137th daya of the year 2005. Figure Shown below is the Magneticfield data between those days and the events like Bow Shock and Magnetopause are marked with blue and red color respectively.

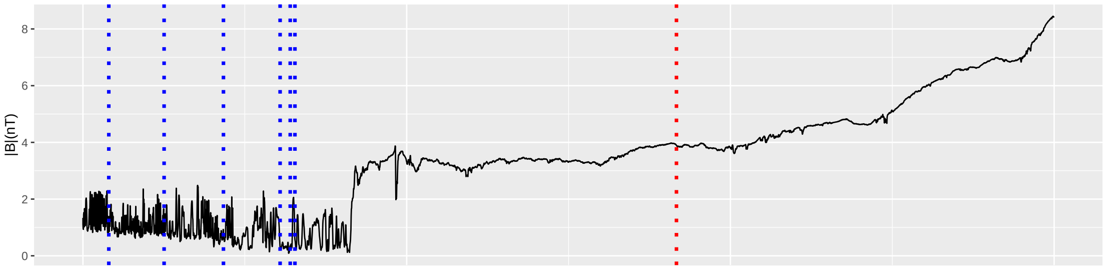

Figure 3: Line Graph of Total Magnetic Field Strength between 136 and 138th day of the year 2005

From the above figure we can clearly see that first boundary crossed by the spacecraft was Bow shock and later followed by the Magnetopause. There are very large fluctuations during when it crossed the Bow Shock boundary but in the case of Magnetopause region the total magnetic field strength was kind of constant.

## 5.1 Trajectory of Spacecraft

It is crucial to analyze the effect of Positions of Spacecraft in predicting the events based on this dataset So, that we can understand the significance of that variable.

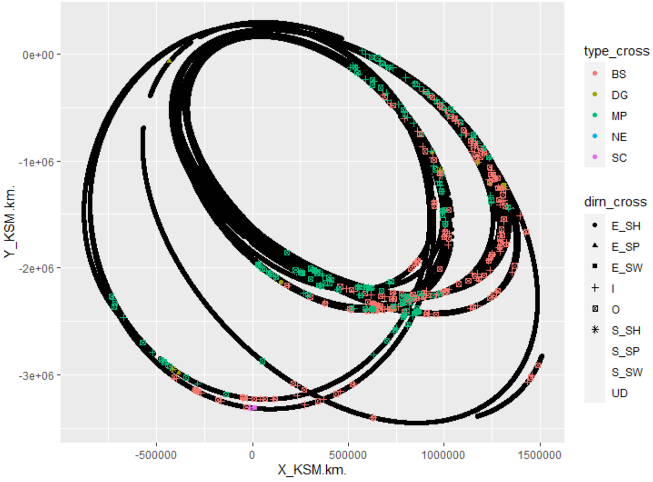

Figure 4: Boundary Crossing points on the Trajectory of Cassini Space craft

From the above plot we can clearly say that the Bow Shock events are in the Outermost orbit on which the spacecraft traveled and all the Magnetopause boundary crossings are in the tail of those orbits. Most of the Bow Shock boundary crossing events has high `X_KSM.km.` value which is greater than 500000 when compared to other events. The `Y_KSM.km.` value for Bow Shock events mostly lies in between 0 and $-2 \times 10^{-3}$. The plot also gives information about the range of `X_KSM.km.` and `Y_KSM.km.` when 90% of boundary crossing happened which is $(0, 11 \times 10^5)$ and $(-2 \times 10^6, 0)$.

## 5.2 Examining the Data Imbalance

It is always important to check whether the given data is balanced or not. If it is not balanced that means there is a large number of particular class and fewer data points for all other classes of data. This imbalance will makes the model that we wants to train biased to that set of data.

```
## .
##     BS     DG     MP     NE     SC
##    245     15    203 494127      2
```

Table 1: Number of Datapoints in each Boundary Class

| Event Class | Count |
|---|---|
| SCAS (unreliable) | 2 |
| Data Gap | 15 |
| Magnetopause | 203 |
| Bow Shock | 245 |
| No Events Occurred | 494127 |
| Total | 494592 |

From the above table, it is clear that 90% of the data points are in the No Events Occurred Class and there are only very few data in the Bowshock and Magnetopause class. So a method must be adopted to prepare the training dataset for model training. Since there are only a very few data points which are in DG and SCAS class So, we can remove those data points because both classes gives unreliable data.

## 5.3 Proportion of Direction of Cross in Different Classes

The `dirn_type` variable explains at which direction the spacecraft was moving when it took the measurements. To get more insights on which direction did the most boundary crossings were recorded.

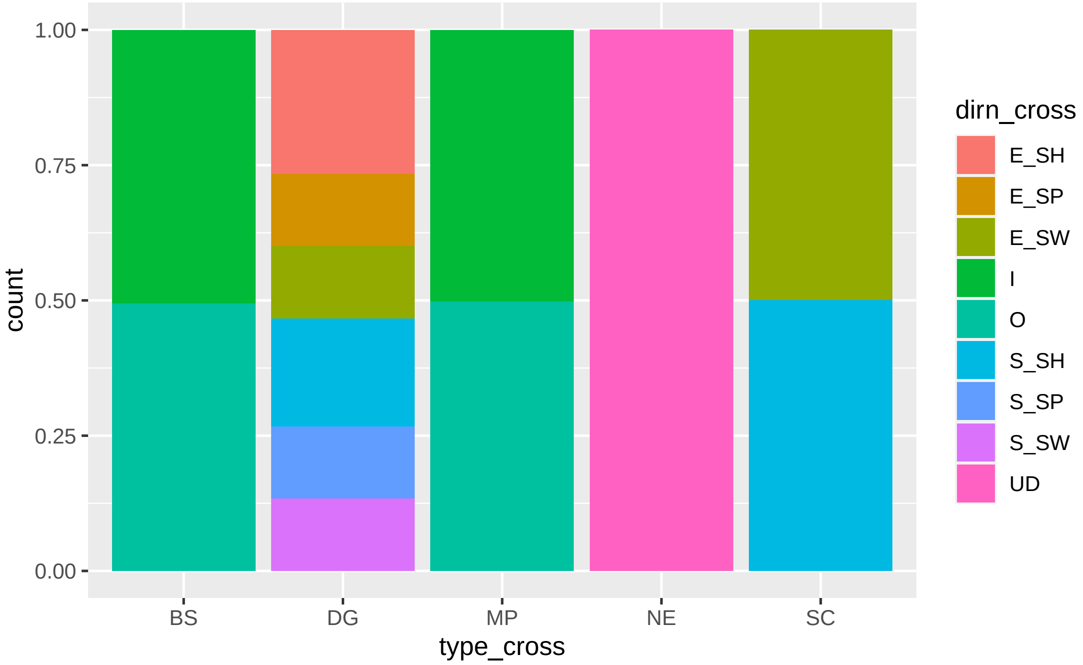

Figure 5: Proportion of different directions on each Class

From this stacked Bar plot, we can see that 50% of both the Bow Shock and Magnetopause events are recorded during Inbound and remaining 50% during the Outbound of the Spacecraft. But for all No events Occurred class the direction of cross is Unknown. All other directions like starting from Magnetosheath, Starting from solar wind, starting from Magnetosphere etc are in SCAS and Datagap class. All datapoints with SCAS class are unreliable data and SCAS data was recorded during when the spacecraft starts with Magnetosheath and Ends with Solar wind. All the data points that were recorded when spacecraft was in the End of Magnetosheath, Starts with solar wind, End of Magnetosphere, End of Magnetosphere and starts with Magnetosphere are in Data Gap class. Since the direction of cross of all `NE` data points are Unknown or not available So, it is better to remove `dirn_cross`.

# 6 Significance of Predictors

Before Model development Significance of each variable must be evaluated so, that we can remove the un necessary variables which will leads to over fitting of models. To find the variable I have used two Logistic Regression Models. Logistic regression Models can be fit to the data by using the maximum likelihood technique. The `family` of the models used was set to `binomial` because we are

trying to classify two classes.

**Train and Test Dataset**

Since the Dataset is a highly imbalanced one with 90% of the data has the class `No Events Occurred`. It is crucial to train the model with a balanced dataset so, for that we sampled 100 datapoints of each class by using `sample_n()` function for Dataset 4 and 300 `NE` Points were chosen instead of 100 for the model with Dataset 3 because If predictor count is nearly as big as total data points then the linear regression is too flexible and overfits the data.All other datapoints were chosen as Test Dataset for this model.

**Standardizing the Train and Test Dataset**

Standardizing is a techniques used in the Machine Learning. The main aim of Normalization is to make all the numeric columns in the dataset to a common scale.All the numeric variables in the train dataset was Normalized by using the `scale()` function. The standard deviation and mean of each variable of train dataset was stored separately. The test dataset was then Standardized by using the standard deviation and Means that were stored before for each variable. For Standardizing I have used the below formula:

$$X' = \frac{(X - \mu)}{\sigma} \tag{2}$$

In this formula mu represents the mean and sigma represents the standard deviation of that variable in the train dataset. By standardizing features we are centering the datapoints to zero and making the standard deviation of value 1.

## 6.1 Logistic Regression Model: Bow Shock vs Other Events

In this model I tried to predict the Bow shock events by using different derived datasets. Since the datasets are highly imbalanced It is important to sample equal number of classes from datasets to train the Logistic Regression model. Before splitting into training and Test Dataset, A new variable called `event_occured` was created and all Bow Shock events were stored as 1 and all the rest of the events as 0 in the `events_occured` variable.

### 6.1.1 Results

Models were trained with different derived datasets and later compared the results.Table 2 Shown below contains the information regarding the datasets used and count of Variables that are significant which is extracted from the summary of the model. From the results of Models that were trained with different datasets, Table 2 indicates that all the predictor variables are significant

Table 2: Logistic Regression : Bow Shock Vs Other Events Results Table

| Dataset Used | Significant Predictors | Total Predictors |
|---|---|---|
| Dataset 3 | 218 | 218 |
| Dataset 4 | 25 | 25 |
| Dataset 3 without Lead Variables | 107 | 107 |

Eventhough the all the predictor variables are significant for all the datasets when it comes to accuracy and recall value each model gives different values.

**Summary of Dataset 3**

From Dataset 3 we have sampled 300 NE points, 100 Bow shock data points and 100 Magnetopause Data points for the train data because there are 218 predictors in this dataset and If predictor count is nearly as big as total data points then the linear regression is too flexible and overfits the data. The Table shown below is the confusion matrix.

```
## .
##  BS  MP  NE
## 100 100 300
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction      0      1
##          0 385553    101
##          1 108361     44
##
##                Accuracy : 0.7805
##                  95% CI : (0.7793, 0.7816)
##     No Information Rate : 0.9997
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 2e-04
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 3.034e-01
##             Specificity : 7.806e-01
##          Pos Pred Value : 4.059e-04
##          Neg Pred Value : 9.997e-01
##              Prevalence : 2.935e-04
##          Detection Rate : 8.906e-05
##    Detection Prevalence : 2.194e-01
##       Balanced Accuracy : 5.420e-01
##
##        'Positive' Class : 1
##
```

From the matrix we can clearly see that the accuracy is about 78% . Eventhough it gives better accuracy the sensitivity of the Model is very low. Accuracy is the percentage of datapoints that are classified correctly but in the case of the Sensitivity which is number of exact positive predictions divided by the total number of positive in this model the sensitivity gives a value that explains about Bow Shock events that are correctly classified . 30.3 is the sensitivity of the above model So, 30.3 out of 100 Bow Shock events were classified correctly through this model.

**Summary of Dataset 4 : Average and Standard Deviation Data**

In the case of this dataset there are only 24 predictors so, the train dataset contains 100 NE points, 100 BS Points and 100 MP points which was sampled from the Dataset 4. The table Shown below is the Confusion matrix of the model with this dataset.

```
## .
##  BS  MP  NE
## 100 100 100
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction      0      1
##          0 484916     87
##          1   9198     58
##
##                Accuracy : 0.9812
##                  95% CI : (0.9808, 0.9816)
##     No Information Rate : 0.9997
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0118
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.4000000
##             Specificity : 0.9813849
##          Pos Pred Value : 0.0062662
##          Neg Pred Value : 0.9998206
##              Prevalence : 0.0002934
##          Detection Rate : 0.0001173
##    Detection Prevalence : 0.0187270
##       Balanced Accuracy : 0.6906924
##
##        'Positive' Class : 1
##
```

From the matrix we can clearly see that the accuracy is about 98.1% . Eventhough it gives better accuracy than the Model with Dataset 3 but still the sensitivity of the Model is similar to that of the Dataset 3 Logistic Regression Model. 40% of the Bow Shock boundaries were classified correctly. But when compared to Model that was trained with Dataset 3, The `NE` points were classified more accurately in this model.The specificity of this model is 98.1% which means the 98.1% of the NE Points were classified correctly through this model.

**Summary of Dataset 3 : Without Lead Variables**

In this dataset there are only 107 predictors which means half of the variables are removed so, the train dataset contains 100 NE points, 100 BS Points and 100 MP points which was sampled from the Dataset 4. The table Shown below is the Confusion matrix of the model with this dataset.

```
## .
##  BS  MP  NE
## 100 100 100
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction      0      1
##          0 448151     68
##          1  45979     77
##
##                Accuracy : 0.9068
##                  95% CI : (0.906, 0.9076)
##     No Information Rate : 0.9997
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0027
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.5310345
##             Specificity : 0.9069496
##          Pos Pred Value : 0.0016719
##          Neg Pred Value : 0.9998483
##              Prevalence : 0.0002934
##          Detection Rate : 0.0001558
##    Detection Prevalence : 0.0931789
##       Balanced Accuracy : 0.7189920
##
##        'Positive' Class : 1
##
```

From the matrix we can clearly see that the accuracy is about 90.7% . It gives better accuracy and sensitivity than the Model with the Dataset 3 Logistic Regression Model. 53.1% of the Bow Shock boundaries were classified correctly. But when compared to Model that was trained with Dataset 3, All other Classes except Bow Shock were classified more accurately in this model.The specificity of this model is 90.7% which means the 90.7% of the NE Points were classified correctly through this model.

## 6.2 Logistic Regression Model: Magnetopause vs Other Events

Unlike the previous model which classifies Bow Shock events and All other Events, This model classify Magnetopause and all other events by using different derived datasets. Since the datasets are highly imbalanced It is important to sample equal number of classes from datasets to train the Logistic Regression model. Before splitting into training and Test Dataset, A new variable called `event_occured` was created and all Magnetopause events were stored as 1 and all the rest of the events as 0 in that newly created variable.

### 6.2.1 Results

Models were trained with different derived datasets and later compared the results.Table Shown below contains the information regarding the datasets used and count of Variables that are significant. A brief summary of models trained with different dataset was explained in the coming pages.

Table 3: Logistic Regression : Magnetopause Vs Other Events Results Table

| Dataset Used | Significant Predictors | Total Predictors |
|---|---|---|
| Dataset 3 | 218 | 218 |
| Dataset 4 | 25 | 25 |
| Dataset 3 without Lead Variables | 107 | 107 |

**Summary of Dataset 3**

From Dataset 3 I have sampled 300 NE points, 100 Bow shock data points and 100 Magnetopause Data points for the train data because there are 218 predictors in this dataset and If predictor count is nearly as big as total data points then the linear regression is too flexible and overfits the data.The Table shown below is the confusion matrix.

```
## .
##  BS  MP  NE
## 100 100 300
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction      0      1
##          0 351892     70
##          1 142064     33
##
##                Accuracy : 0.7123
##                  95% CI : (0.711, 0.7136)
##     No Information Rate : 0.9998
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 3.204e-01
##             Specificity : 7.124e-01
##          Pos Pred Value : 2.322e-04
##          Neg Pred Value : 9.998e-01
##              Prevalence : 2.085e-04
##          Detection Rate : 6.679e-05
##    Detection Prevalence : 2.876e-01
##       Balanced Accuracy : 5.164e-01
##
##        'Positive' Class : 1
##
```

From the matrix we can clearly see that the accuracy is about 71.2% . Eventhough it gives a better accuracy value the sensitivity of the Model is very low. Accuracy is the percentage of datapoints that are classified correctly but in the case of the Sensitivity which is number of exact positive predictions divided by the total number of positive in this model the sensitivity gives a value that explains about Bow Shock events that are correctly classified . 32% is the sensitivity of the above model So, 32 out of 100 Bow Shock events were classified correctly through this model. When Comparing with the Logistic Regression Model that was used for Classifying the Bow Shock and Rest of the Events, this Logistic regression model was able to classify the Magnetopause events more efficiently by using Dataset 3.

**Summary of Dataset 4 : Average and Standard Deviation Data**

In the case of this dataset there are only 24 predictors so, the train dataset can sample 100 NE points, 100 BS Points and 100 MP points which is sampled from the Dataset 4. The table Shown below is the Confusion matrix of the model with this dataset.

```
## .
##  BS  MP  NE
## 100 100 100
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction      0      1
##          0 490008     58
##          1   4148     45
##
##                Accuracy : 0.9915
##                  95% CI : (0.9912, 0.9917)
##     No Information Rate : 0.9998
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0206
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 4.369e-01
##             Specificity : 9.916e-01
##          Pos Pred Value : 1.073e-02
##          Neg Pred Value : 9.999e-01
##              Prevalence : 2.084e-04
##          Detection Rate : 9.105e-05
##    Detection Prevalence : 8.483e-03
##       Balanced Accuracy : 7.142e-01
##
##        'Positive' Class : 1
##
```

From the matrix we can clearly see that the accuracy is about 99.1% . Eventhough it gives better accuracy than the Model above with Dataset 3 but still the sensitivity of the Model is similar to that of Logistic Regression Model which classifies Magnetopause Events by using the Dataset3. 43.7% of the Bow Shock boundaries were classified correctly. But when compared to Model that was trained with Dataset 3, The `NE` points were classified more accurately in this model.The specificity of this model is 99.2% which means the 99.2% of the NE Points were classified correctly through this model.

**Summary of Dataset 3 : Without Lead Variables**

In this dataset there are only 107 predictors which means half of the varibales are removed so, the train dataset contains 100 NE points, 100 BS Points and 100 MP points from the Dataset 4. The table Shown below is the Confusion matrix of the model with this dataset.

```
## .
##  BS  MP  NE
## 100 100 100
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction      0      1
##          0 432492     50
##          1  61680     53
##
##                Accuracy : 0.8751
##                  95% CI : (0.8742, 0.876)
##     No Information Rate : 0.9998
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0013
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.5145631
##             Specificity : 0.8751852
##          Pos Pred Value : 0.0008585
##          Neg Pred Value : 0.9998844
##              Prevalence : 0.0002084
##          Detection Rate : 0.0001072
##    Detection Prevalence : 0.1248961
##       Balanced Accuracy : 0.6948741
```

```
##
##        'Positive' Class : 1
##
```

From the matrix we can clearly see that the accuracy is about 87.5% . It gives better accuracy and sensitivity than the Model with the Dataset 3 Logistic Regression Model. 51.5% of the Bow Shock boundaries were classified correctly. But when compared to Model that was trained with Dataset 3, All other Classes except Bow Shock were classified more accurately in this model.The specificity of this model is 87.5% which means the 87.5% of the NE Points were classified correctly through this model.

# 7 Classification Model Development

This section explains about the Model that was Developed for the classification of different Boundaries based on the Magnetometer data and the analysis that were done before. In the Predictor Significance section we found that Dataset 4(Average Dataset) performed better when compared to the other datasets. In this section we will compare the results of Models trained with different datasets and conclude which model can be used to address the problem statement. The model with high sensitivity value and accuracy will be considered as the Best Model. Because sensitivity value of each class will tell us the percentage of each class that was correctly classified.

## 7.1 Random Forest Model

This section explains about the three Random forest Models that were trained by using three different datasets. The model will be selected based on lowest classification error rate especially when it comes to Magnetopause and Bow Shock event Classification. Several Methods can be used for classifying the boundaries by using these kind of datasets, but in this project we are going to use random forest model. Random Forest Model is a Supervised Learning algorithm where it builds an ensemble of decision trees. Random forest produces great results by handling large datasets with higher dimensionality. Additional randomness will be added to the model by the random Forest algorithm, while growing the trees. It looks for the best predictor among a random set of predictors which will generally results in a better model.

### 7.1.1 Setting up Random Forest Model

In this project i have used the `ranger()` package because it provides a faster implementation of random forest and also easy to tune the hyper parameters.Here the model uses following configuration:

- `num.trees` was set to 500. `num_trees` will determine how many trees the algorithm builds before it takes the voting or averages of the prediction. Higher the number of trees will usually gives better performance but makes the computation a lot slower.
- `verbose` which shows the computation status and Estimated runtime was set to TRUE
- `importance` was set to `impurity` which will give the variable importance. impurity measures the Gini Index for the classification.

### 7.1.2 Train and Test Dataset

Like we discussed in Section 5 the dataset is a highly imbalanced one with 90% of the data has the class `No Events Occurred`. It is important to train the model with a balanced dataset so, for that we sampled 100 datapoints of each class by using `sample_n()` function for Dataset 4.Both the train and test dataset was standardized by using the same technique that was discussed in the section 6.

### 7.1.3 Results

**Summary of Dataset 3**

From Dataset 3 I have sampled 300 NE points, 100 Bow shock data points and 100 Magnetopause Data points for the train data because there are 218 predictors in this dataset and If predictor count is nearly as big as total data points then the linear regression is too flexible and overfits the data.The Table shown below is the confusion matrix.

```
## .
##  BS  MP  NE
## 100 100 300
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     BS     MP     NE
##         BS     24      3  35997
##         MP      6     44  49709
##         NE    115     56 408105
##
## Overall Statistics
##
##                Accuracy : 0.8262
##                  95% CI : (0.8251, 0.8272)
##     No Information Rate : 0.9995
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 9e-04
##
##  Mcnemar's Test P-Value : <2e-16
##
## Statistics by Class:
##
##                      Class: BS Class: MP Class: NE
## Sensitivity          1.655e-01 4.272e-01 0.8264397
## Specificity          9.271e-01 8.994e-01 0.3104839
## Pos Pred Value       6.662e-04 8.843e-04 0.9995812
## Neg Pred Value       9.997e-01 9.999e-01 0.0008976
## Prevalence           2.935e-04 2.085e-04 0.9994980
## Detection Rate       4.858e-05 8.906e-05 0.8260248
## Detection Prevalence 7.291e-02 1.007e-01 0.8263709
## Balanced Accuracy    5.463e-01 6.633e-01 0.5684618
```

### Summary of Dataset 4 : Average and Standard Deviation Data

In the case of this dataset there are only 24 predictors so, the train dataset can sample 100 NE points, 100 BS Points and 100 MP points which is sampled from the Dataset 4. The table Shown below is the Confusion matrix of the model with this dataset.

```
## .
##  BS  MP  NE
## 100 100 100
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     BS     MP     NE
##         BS    129     17      0
##          MP     16     86      0
##          NE      0      0 494011
##
## Overall Statistics
##
##                Accuracy : 0.9999
##                  95% CI : (0.9999, 1)
##     No Information Rate : 0.9995
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9334
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: BS Class: MP Class: NE
## Sensitivity          0.8896552 0.8349515    1.0000
## Specificity          0.9999656 0.9999676    1.0000
## Pos Pred Value       0.8835616 0.8431373    1.0000
## Neg Pred Value       0.9999676 0.9999656    1.0000
## Prevalence           0.0002934 0.0002084    0.9995
## Detection Rate       0.0002610 0.0001740    0.9995
## Detection Prevalence 0.0002954 0.0002064    0.9995
## Balanced Accuracy    0.9448104 0.9174595    1.0000
```

From the matrix we can clearly see that the accuracy is about 87.5% . Eventhough it gives better accuracy than the Model above with Dataset 3 but still the sensitivity of the Model is similar to that of Logistic Regression Model which classifies Magnetopause Events by using the Dataset3. 51.5% of the Bow Shock boundaries were classified correctly. But when compared to Model that was trained with Dataset 3, The `NE` points were classified more accurately in this model.The specificity of this model is 87.5% which means the 87.5% of the NE Points were classified correctly through this model.

# 8 References

Sergis, Jackman, N. 2013. "Particle and Magnetic Field Properties of the Saturnian Magnetosheath: Presence and Upstream Escape of Hot Magnetospheric Plasma." *Journal of Geophysical Research: Space Physics*, nos. 118, 1620–1634. https://doi.org/10.1002/jgra.50164.

# 9 Appendix

## 9.1 Supporting code