

# Churn Propensity Project

## Roles and responsibilities

Business Stakeholder: [Duijn, Albert van \(ELS-AMS\)](#)

Analyst: [Daniel Momoh \(ELS-LOW\)](#)

Baringa:

Lead Specialist: Siddesh Dhuri

Analyst: Alex Leigh

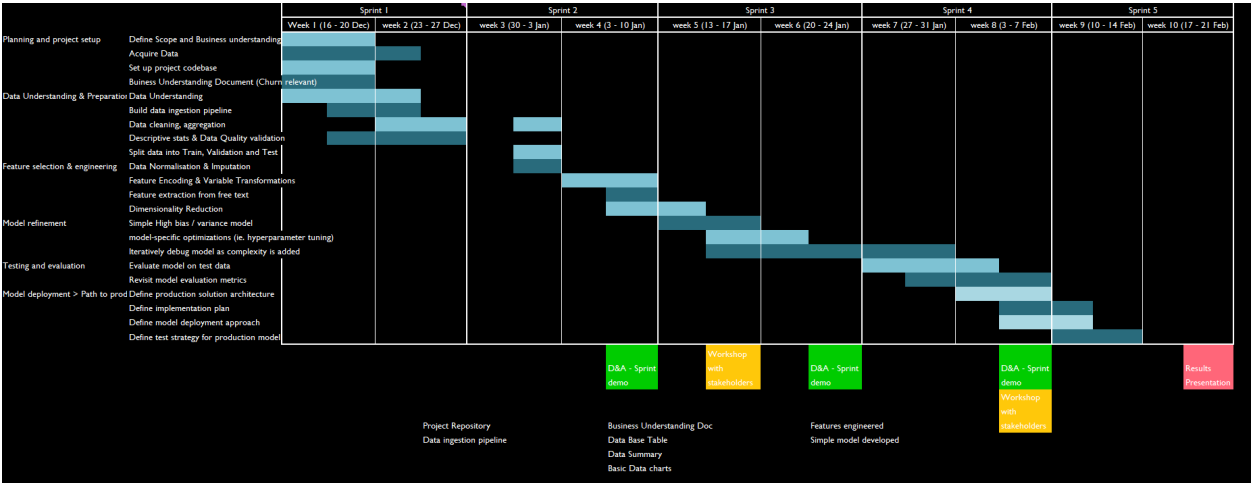
## Preamble

The output of this Project will be a Machine Learning model built in Python that ingests business data (churn, cancellations, product, customer, interactions, NPS etc.) and predicts the likelihood of an Elsevier customer to cancel their subscription. The steps that would be required to take the model into production will also be defined. The model will work at product line level 2.

## Project Plan

This project will be approached following the CRISP-DM methodology, focusing efforts on 5 sprints which achieve intermediate goals and produce a series of deliverables. The plan is outlined below and covers the following key stages:

- 1. Planning and Project Setup
- 2. Data Understanding and Preparation
- 3. Feature Selection and Engineering
- 4. Model Refinement
- 5. Testing and Evaluation
- 6. Model Deployment > Path to Production



Alongside working directly on the model, we will hold a series of demos which will look at what has been achieved so far, and there will also be two workshops with key stakeholders aimed at maintaining alignment with business expectations.

We will be using a Jira Kanban board for agile project management. The Jira board can be found at <https://jira.cbsels.com/secure/RapidBoard.jspa?projectKey=RSSC&rapidView=3319>

## Project setup:

Although we have leveraged the CRISP-DM methodology to structure the project. To ensure that we can accommodate the data received late, new data sources identified, changes in definition of churn, and other vagaries of project, we also employed agile methodology and tools to structure project works and drive collaboration.

We have had multiple workshops with Data & Analytics team, Sales enablement teams and Sales leaders to capture their feedback.

We used JIRA for managing project in an agile manner.

A project management tool such as JIRA helps to improve visibility of tasks in the backlog, those selected for development, in progress and completed. Information such as task owner, task co-dependencies and which epic a task is linked to is readily available for everyone. On an agile project, Jira Kanban boards can be used as the basis for discussion during daily stand-ups.



### Kanban board

QUICK FILTERS: [Only My Issues](#) [Recently Updated](#)

14 Backlog

5 Selected for Development

<p><b>RSSC-1</b></p> <p>Developing an understanding of the sales and marketing data received from context of churn propensity modelling and preparing data for modelling</p>	<p><b>RSSC-8</b></p> <p>As an Engineer, I want to aggregate all data tables into one base table, so that i can start model development</p> <p><b>Data Understanding &amp; Preparation</b></p>
<p><b>RSSC-2</b></p> <p>Parsing and Pruning data to identify and engineer relevant features</p>	<p><b>RSSC-35</b></p> <p>As an engineer, I want to maintain the confluence and produce a deck summarising our approach to this project so that we can clearly communicate the project to business stakeholders</p> <p><b>Path to Production</b></p>
<p><b>RSSC-3</b></p> <p>Developing machine learning model for predicting customer product churn propensity</p>	<p><b>RSSC-36</b></p> <p>As an engineer, I want to ascertain the product level in every table that has products, so that we can map to product level 2 correctly</p> <p><b>Feature selection &amp; engineering</b></p>
<p><b>RSSC-4</b></p> <p>Evaluating model performance of test and validation data</p>	<p><b>RSSC-48</b></p> <p>As a Business Stakeholder, I want to understand metric related to churn, so that i can understand the current status of business</p>
<p><b>RSSC-19</b></p> <p>Define and outline a path to production for the ML model</p>	<p><b>RSSC-52</b></p> <p>As a business stakeholder, I want to see a baseline model and results in the workshop so that i can understand the process and outputs to expect in mode detail.</p> <p><b>Model development and refinement</b></p>
<p><b>RSSC-17</b></p> <p>As an business stakeholder, I want to have one usage metric for most of the products, so that i have as many products as possible with usage metric</p>	

The project repository is set up in GitLab, you can follow the link : <https://gitlab.et-scm.com/business-analytics/rss-customer-churn/tree/master>

Name	Last commit
data	Model development and churn analysis jupyter notebooks
jupyter	Model development and tuning workbooks committed
...	...

src	model development and churn analysis jupyter notebooks
.gitignore	RSSC34: code to create HDF file in HDF data store for all except...
Pipfile	XGBoost model experiment added
Pipfile.lock	XGBoost model experiment added
README.md	added command to execute
config.yaml	Model development after new definition of churn
README.md	

A Git usage training session was also conducted and documentation created to get started using git to help set up git repository to encourage collaboration and future branch developments

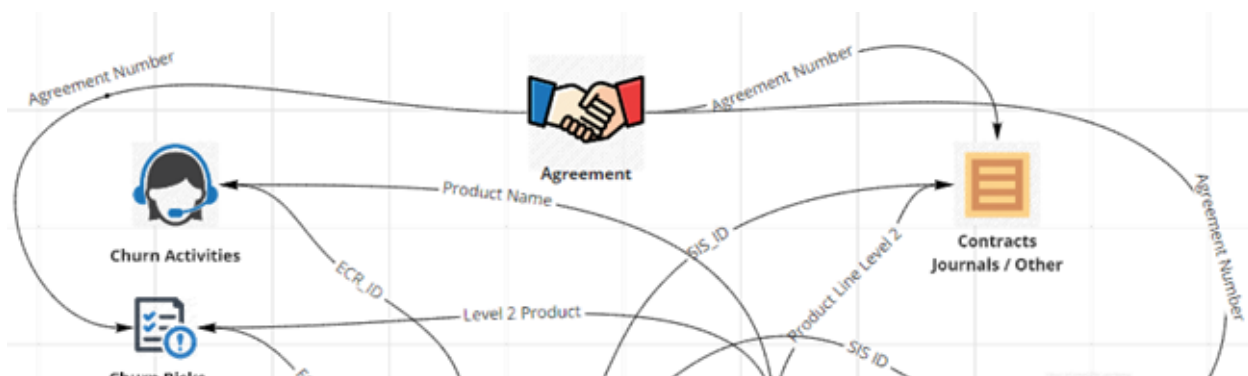
## Data Acquisition and Understanding

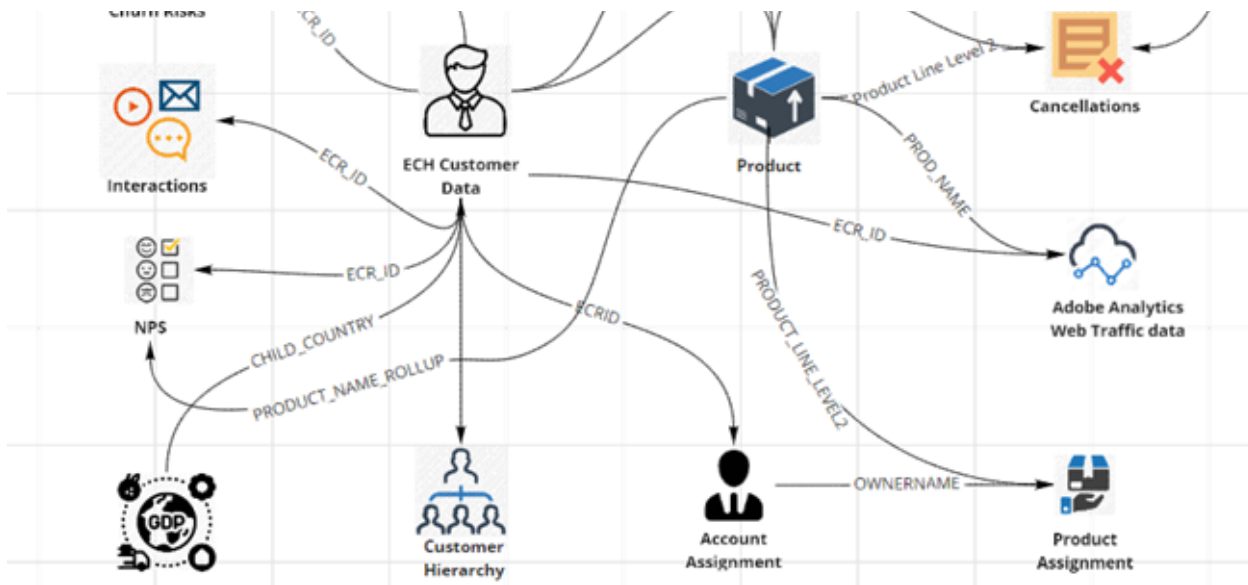
- Data acquired
  - We received data from the business in .xlsx, .csv and .gz form. this data was stored in the shared directory: /Reed Elsevier Group ICO Reed Elsevier Inc/Duijn, Albert van (ELS-AMS) - Churn Propensity Modelling/02. Data collection
  - Some files for example GEDR data, Hierarchy data were shared in email,
  - To ensure that all data can be located in one place, but also not to disturb the shared location, the data is available in the project folder as well.
  - Some files were very large, and we have ingested these into python native HDF and Pickle formats for faster processing
- Data understanding developed
  - Key to a Machine Learning project is maintaining consistent communication with the business.
  - Data mapping is maintained in the shared folder at location: Reed Elsevier Group ICO Reed Elsevier Inc/Duijn, Albert van (ELS-AMS) - Churn Propensity Modelling/data\_mapping.xlsx

191212_churn_activities	191224_churn_risks_V02	DataCR_from_2015_Cancellations	DataCR_from_2015_journals	DataCR_from_2015_other_products	ECH Customer Data
Opportunity	Opportunity ID	SIS Id (Agreement SIS)	SIS Id (Agreement SIS)	SIS Id (Agreement SIS)	ecrid
Created By	Opportunity Name	HQ SIS Id (Agreement SIS)	HQ SIS Id (Agreement SIS)	HQ SIS Id (Agreement SIS)	name
Account ID	Sales Type	Name (Agreement SIS)	Name (Agreement SIS)	Name (Agreement SIS)	city
Company / Account	Stage	Business Division (Agreement SIS)	Business Division (Agreement SIS)	Business Division (Agreement SIS)	Country ISO
Contact	Level 1 Product	Sales Division (Agreement SIS)	Sales Division (Agreement SIS)	Sales Division (Agreement SIS)	Region
Lead	Level 2 Product	Division	Division	Division	consortium
Priority	Expected Close Date	RSO	RSO	RSO	post_code
Activity Type	Subscription Start Date	Subregion Grouping	Subregion Grouping	Subregion Grouping	Classification
Task	Amount (converted) Currency	Country Name (Agreement SIS)	Country Name (Agreement SIS)	Country Name (Agreement SIS)	
Task/Event Record Type	Amount (converted)	WIP Flag	WIP Flag	WIP Flag	
Task Subtype	Agreement Number	Status	Status	Status	
Event Subtype	Account Name: ECR Id	Wip Type			
Subject	Account Name: Account	Business Indicator	Business Indicator	Business Indicator	

	Name				
Call Result	Risk ID	Sales Type	Sales Type	Sales Type	
Topics Discussed	Risk Name	Calculated New/Renewal	Calculated New/Renewal	Calculated New/Renewal	
Comments	Risk Type		Invoice Num	Invoice Num	
Full Comments	Severity		Invoice Date	Invoice Date	
Follow Up Subject	Status		Payment Term	Payment Term	
Follow Up Notes	Created Date		Payment Term Description	Payment Term Description	
Name of Value Prop	Comments		Payment Term Type	Payment Term Type	
Activity ID	Competitor: Account Name	Status Change Date	Status Change Date	Status Change Date	
Assigned		Reporting Year (2015)	Renewal Exp Complete Date	Renewal Exp Complete Date	
Date		Renewal Exp Complete Date	Product Revenue Type	Product Revenue Type	
Product Name		Product Revenue Type	Product Line Level 1	Product Line Level 1	
Assigned Role		Product Line Level 1	Product Line Level 2	Product Line Level 2	
Assigned Role Display		Product Line Level 2	Product Line Level 3	Product Line Level 3	
Created Date		Product Line Level 3	Product Line Level 4	Product Line Level 4	
Start		Product Line Level 4	Saleable Product Name (Source)	Saleable Product Name (Source)	
End		Saleable Product Name (Source)			
ECR ID		Agreement Number	Agreement Number	Agreement Number	
Parent ECR-ID		Agreement Start Date	Agreement Start Date	Agreement Start Date	
Parent ECR-ID		Agreement End Date	Agreement End Date	Agreement End Date	
		Subscription Start Date	Subscription Start Date	Subscription Start Date	
		Subscription End Date	Subscription End Date	Subscription End Date	
		Parent Agreement Number	Parent Agreement Number	Parent Agreement Number	
		Bookings - Final Net Price - Agent Discount Amount(Rep) \$	Bookings - Final Net Price - Agent Discount Amount (Rep)	Bookings - Final Net Price - Agent Discount Amount (Rep)	
		Bookigns - Committed Print(Rep) \$	Bookigns - Committed Print (Rep)	Bookigns - Committed Print(Rep)	
		Cancellation Month			
		Cancellation Reason			

In order to visualise links between different tables, we have developed the graph shown below. The output of this analysis will be the basis of our work in joining the tables provided to us. The nodes represent files, and on each arrow, we have specified the linking field name as it is displayed.





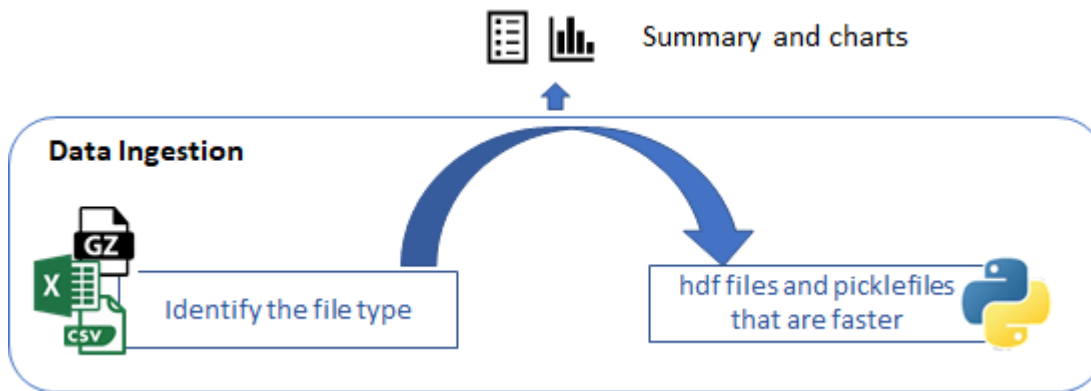
In a Machine Learning project like this one, understanding the business is paramount to producing a model that delivers tangible business value. Upon landing, we had a series of meetings with the business to understand the data sources provided to us, as well as their position in the wider business context.

The data sources acquired are summarized below and can also be found here: <C:\Users\<username>\Reed Elsevier Group ICO Reed Elsevier Inc\Duijn, Albert van (ELS-AMS) - Churn Propensity Modelling\02. Data collection>

File	Data Owner
191224_churn_risks_V02	Anton Jumelet
191212_churn_activities	Anton Jumelet
interaction	Georgia Durston
DataCR_from_2015_Cancellations	Joanna Aksiuto
DataCR_from_2015_journals	Joanna Aksiuto
DataCR_from_2015_other_products	Joanna Aksiuto
ECH Customer Data	Marc Hansen
NPS_Cleansed_Data	Georgia Durston
Hierarchy Data	Artur
GEDR, Data	Joanna
Web Traffic Data	Harry Wilkes
SD Usage Data (Artur Extract)	Harry Wilkes / Artur
Account Assignment	Anton
Product Assignment	Anton

## Data Ingestion

We initially ingested the datasets by saving them as pickle files; these are more easily stored and used in subsequent code. Unfortunately, some of the datasets are quite large and handling them we would run into memory issues, and pickle file can perform badly with large datasets. as a result we moved most of the data into HDF store. this can be found in projects data/hdf folder/



Data can be ingested from the xls / csv / gz file in the data directory and into the Python native formats that are used for machine learning using the data ingestion command line interface as. please ensure you have enough memory before running data ingestion. data ingestion can run for 20-30 mins.

to only ingest data

```
PS C:\Users\DHURIS\REPO\rss-customer-churn> python src/main.py ingest
```

to ingest and summarise data

```
PS C:\Users\DHURIS\REPO\rss-customer-churn> python src/main.py ingest summarise
```

this will create a summary.txt in the project data/summary directory.

## Data Cleaning

Often, data is provided with fields incomplete or completed incorrectly. Cleansing and imputing is essential to creating a full dataset that a model can work on.

### Handling Missing Data

#### Examples of Missing Fields:

| Blank | Null | Nan | N.A. | Unknown |  
| Don't know | 00-00-0000 | £0 |

#### Reasons for Missing Data

People don't know how to fill in the field
Completing the field is too time-consuming or boring
The field is not included as a mandatory field

## Variable Encoding and Transformation

Once the data is clean and full, it must be prepared for the model. Clean data will likely be in a digestible, readable form for humans, but this is not appropriate for a Machine Learning model. We must now translate the data into a form that we can feed into the model.

Although the general steps of cleansing and standardising are formulaic, this process requires human intervention as every dataset is different and is surrounded by business context that a computer would not understand. We have made these steps reproducible in Jupyter Notebooks, but upon the ingestion of a new dataset, small alterations to the code would have to be made to reflect this. E.g. which columns are redundant? How will we impute missing data for each column? This will vary and requires a human to interpret the data.

## 10-Step Data Cleansing and Standardisation Process:

We have developed a series of Jupyter notebooks which make the standardisation process simple to follow and reproducible with future datasets. To summarise the steps we took:

1. Replace any values representing a null value (e.g. 'NA', 'Unknown', 00) with a null
2. Identify what percentage of the column is missing data and remove any columns above the threshold
3. Impute missing values in the remaining columns with the appropriate method (e.g. mean, median, most frequent)
4. Detect and remove outliers
5. Bin the less frequently occurring values into a category such as 'other' to prevent overfitting
6. Log transform any numerical columns with values varying by orders of magnitude
7. Scale columns to create consistent ranges of the data
8. One-hot encode any columns with categorical data
9. Transform any date columns into a standard format
10. Remove any unused or redundant columns

All Jupyter notebooks for this project can be found in `<\rss-customer-churn\jupyter\Cleaning>` in the repository.

## Definition of churn

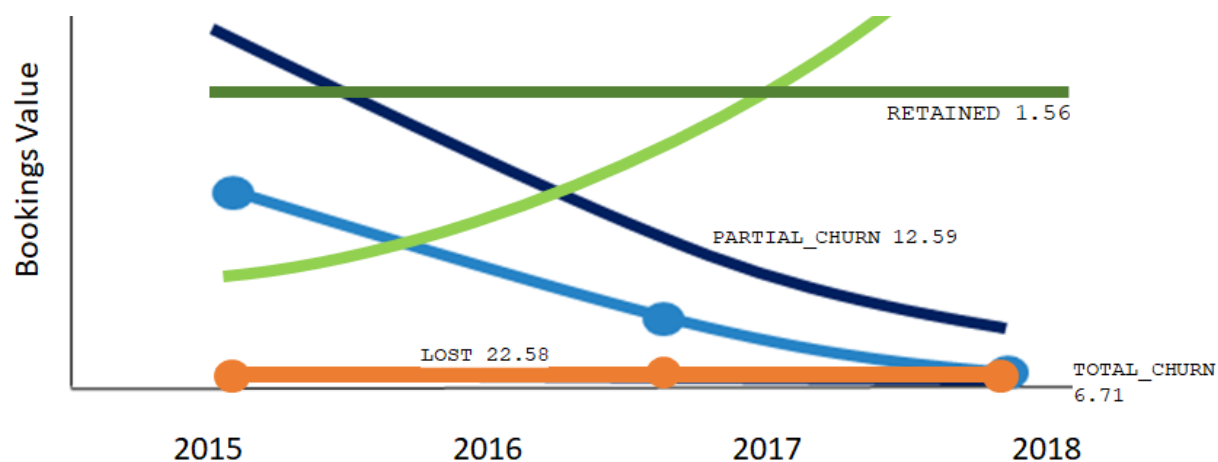
Definition of churn




As we started the project the definition of churn was the cancellation of an agreement. However during our workshop feedback from business suggested that subscription cancellation does not always constitute churn as customers are often upsold on a different subscription, thus triggering cancellation of existing subscription

We define churn into two categories

- Total churn: when the customer value for a product goes down to zero or lower than zero.
- Partial churn: when the customer value is more than zero for a product but is lower than in the previous year.

Here we consider churn at product line level 2. thus a customer might have no churn on a journal product but total churn on SCOPUS and partial churn on Reaxys.



		2015	2016	2017	2018	
		\$100	\$120	\$150	\$160	UPSOLD
		\$100	\$80	\$ 60	\$50	PARTIAL CHURN
		\$100	\$50	\$30	\$0	TOTAL CHURN
		\$100	\$0	\$0	\$0	LOST

#### Distribution of partial churn customers

- We distribute the partial churn customers into categories to determine the size of impact of partial churn.
- We see the in 2018 38% of customers that partially churned lost value in the range 40-50% compared to previous year
- The second largest chunk of 15% is 10-20% value lost over last year.
- However the third largest category of 14% is partial churn of over 80% value lost. These customers are very close to total churn.
- The distribution is fairly consistent over the 3 year period, with the top two categories remaining unchanged.

Category	2018
40 to 50 pct churn	38%
10 to 20 pct churn	15%
over_80_pct_churn	14%
0 to 5 pct churn	13%
5 to 10 pct churn	12%
20 to 30 pct churn	11%
30 to 40 pct churn	10%



50 to 60 pct churn	<div></div>	9%
60 to 70 pct churn	<div></div>	7%
70 to 80 pct churn	<div></div>	6%

## Feature Engineering

During Model development, data scientists create a set of attributes (input features) that represent various behaviour patterns related to customer engagement level with a service or product. In a broad sense, features are measurable characteristics of observations that an ML model takes into account to predict outcomes (in our case the decision relates to churn propensity.)

Feature engineering is the most important phase of model development as these are the predictors that will help you to develop the model.

Based on literature review and previous experience in subscription churn and publishing businesses we had identified features broadly in the 6 classes depicted in the diagram here. Also to gain understanding from the business and sales reps, we conducted a qualitative workshop to get their first hand experience on the factors that might be indicators of customer churn.

- Customer Characteristics
- Product Characteristics
- Transaction History
- Customer Engagement
- User Experience
- Economic indicators

The benefits of feature engineering include:

- Better features means simpler models.
- Summarising numerical data into practical, aggregate metrics such as mean, sums
- Reduce model dimensionality to avoid the “curse of dimensionality”
- Getting data in the correct format for the mathematical model

### Examples of New Features (calculated from existing columns)

total\_bookings (Total Value of customer)

Length\_of\_relationship (from agreement\_start and agreement\_end)

num\_of\_agreements (frequency in the 5 year period)

mean\_time\_spend\_hrs (user web traffic)

max\_score (form score)



Customer  
Characteristics



Product  
Characteristics



Transactions  
History



Customer  
Engagement



User  
Experience



External  
Factors

business intelligence

- Industry
- Size
- Revenue

business intelligence

- Type
- Offline/online

business intelligence

- Recency
- Frequency
- Money
- Length of relationship

business intelligence

- CSAT
- NPS
- Complaints

business intelligence

- Hours Spent
- No. of visits
- Ease of use

business intelligence

- R&D Spend budget
- Economic indicators

Please find below the list of all features engineered. this is also available in the Jupyter notebook model\_development.ipynb and can be further inspected there.

0	total_bookings	[[500708928.91848654, 90079946.57200013, 68692...	7469	0	float64	0.00
49	num_risks_Very Low	[[0.0, 2.0, 1.0]]	3	0	float64	0.00
56	num_nps	[[3.0, 1.0, 2.0, 18.0, 4.0, 6.0, 8.0, 19.0, 5....	27	0	float64	0.00
55	num_owners_x	[[2.0, 12.0, 39.0, 1.0, 17.0, 88.0, 3.0, 6.0, ...	74	0	float64	0.00
54	mean_days_to_initial_response	[[0.0, 0.8707693750000001, 1.231716775868241, ...	3525	0	float64	0.00
53	max_days_to_initial_response	[[0.0, 5.20266, 47.1442, 98.3448, 38.4664, 80....	3520	0	float64	0.00
52	max_days_to_close	[[0.0, 54.8378, 114.882, 170.49200000000002, 4...	3731	0	float64	0.00
51	mean_days_to_close	[[0.0, 7.792073500000001, 2.4605359131313063, ...	3738	0	float64	0.00
50	num_incidents	[[2.0, 23.0, 4490.0, 1.0, 3623.0, 369.0, 18117...	893	0	float64	0.00
48	num_risks_Medium	[[0.0, 1.0, 4.0, 51.0, 2.0, 11.0, 17.0, 23.0, ...	85	0	float64	0.00
58	min_nps	[[1.0, 5.0, 7.0, 8.0, 10.0, 0.0, 6.0, 2.0, 3.0...	11	0	float64	0.00
47	num_risks_Low	[[0.0, 1.0, 31.0, 8.0, 15.0, 32.0, 9.0, 21.0, ...	108	0	float64	0.00
46	num_risks_High	[[1.0, 13.0, 0.0, 182.0, 62.0, 26.0, 17.0, 5.0...	94	0	float64	0.00
45	num_risks_Critical	[[23.0, 0.0, 31.0, 1.0, 12.0, 9.0, 45.0, 16.0,...	77	0	float64	0.00
44	num_activities_Virtual	[[51.0, 0.0, 7.0, 16.0, 114.0, 13.0, 11.0, 3.0...	92	0	float64	0.00
43	num_activities_Phone	[[9.0, 0.0, 7.0, 6.0, 2.0, 1.0, 3.0, 22.0, 4.0...	62	0	float64	0.00
42	num_activities_Other	[[0.0, 2.0, 1.0, 4.0, 3.0, 5.0, 11.0, 7.0, 14....	15	0	float64	0.00
41	num_activities_Online	[[0.0, 2.0, 14.0, 7.0, 1.0, 6.0, 26.0, 21.0, 1...	46	0	float64	0.00
57	mean_nps	[[4.0, 8.0, 5.0, 9.0, 10.0, 3.0, 7.0, 6.0, 2.0...	11	0	float64	0.00
59	max_nps	[[7.0, 10.0, 9.0, 5.0, 8.0, 6.0, 4.0, 2.0, 0.0...	11	0	float64	0.00
1	total_mean_bookings	[[59150.49367022877, 4578.861717684143, 3118.5...	7478	0	float64	0.00
69	TERRITORY_OWNERRAD	[[0.0, 1.0]]	2	0	float64	0.00
76	gerd_trend	[[updown_trend, increasing_trend, decreasing_t...	3	0	object	0.00
75	gerd_2018	[[55710.29935, 158782.8609, 483759.6158, 16914...	60	0	float64	0.00
74	TERRITORY_OWNERSSMCC	[[1.0, 2.0, 4.0, 3.0, 5.0, 0.0]]	6	0	float64	0.00

73	TERRITORY_OWNERSSM	[[7.0, 9.0, 6.0, 3.0, 4.0, 8.0, 5.0, 2.0, 1.0,...	11	0	float64	0.00
72	TERRITORY_OWNERSD	[[1.0, 2.0, 0.0, 3.0]]	4	0	float64	0.00
71	TERRITORY_OWNERRSSD	[[0.0, 2.0, 1.0]]	3	0	float64	0.00
70	TERRITORY_OWNERRM	[[0.0, 1.0, 2.0]]	3	0	float64	0.00
68	TERRITORY_OWNERCMM	[[1.0, 4.0, 0.0, 2.0]]	4	0	float64	0.00
60	SIZE	[[Medium (=Direct), Large (=Key Accounts), Sma...	4	0	object	0.00
67	TERRITORY_OWNERCMD	[[1.0, 2.0, 0.0]]	3	0	float64	0.00
66	TERRITORY_OWNERCCSD	[[1.0, 2.0, 0.0, 3.0]]	4	0	float64	0.00
65	TERRITORY_OWNERCC	[[6.0, 4.0, 5.0, 2.0, 3.0, 1.0, 7.0, 0.0]]	8	0	float64	0.00
64	TERRITORY_OWNERAM	[[1.0, 2.0, 0.0]]	3	0	float64	0.00
63	TERRITORY_OWNERAGENT	[[0.0, 1.0]]	2	0	float64	0.00
62	num_owners_y	[[19.0, 26.0, 18.0, 14.0, 15.0, 25.0, 12.0, 13...	21	0	float64	0.00
61	TIER	[[T4 - EF, T2 - RE, T1 - RF, T3 - ER]]	4	0	object	0.00
40	num_activities_Face to Face	[[6.0, 10.0, 36.0, 34.0, 9.0, 205.0, 7.0, 22.0...	96	0	float64	0.00
39	num_activities_Email	[[1.0, 0.0, 15.0, 6.0, 12.0, 2.0, 5.0, 3.0, 10...	19	0	float64	0.00
38	users_3_year_change	[[0.0]]	1	0	float64	0.00
10	prod_mean_bookings	[[59487.459545099824, 4650.958946587704, 3129....	10429	0	float64	0.00
17	Classification	[[Academic, Government, Unknown, Corporate, Me...	7	0	object	0.00
16	prod_bookings_per_year	[[83302272.51632145, 29054540.53933339, 972108...	10418	0	float64	0.00
15	prod_length_of_relationship	[[6, 3, 4, 7, 1, 5, 2, 8, 9]]	9	0	int64	0.00
14	prod_days_since_first_agreement	[[2246, 1881, 1516, 2611, 2977, 3342, 1150, 37...	255	0	int64	0.00
13	prod_days_since_last_agreement	[[420, 1150, 239, 1881, 116, 785, 2246, 1516, ...	222	0	int64	0.00
12	prod_num_agrmts_with_parents	[[6, 15, 10, 5, 20, 7, 9, 3, 2, 11, 8, 18, 14,...	32	0	int64	0.00
11	prod_num_agrmts	[[15, 17, 11, 9, 10, 20, 8, 13, 14, 6, 7, 26, ...	38	0	int64	0.00
9	prod_bookings	[[499813635.0979287, 87163621.61800016, 291632...	10412	0	float64	0.00
37	user_trend	[[no_traffic_data]]	1	0	object	0.00
8	over_million_year	[[over_million_per_year, below_million_per_year]]	2	0	object	0.00
7	total_bookings_per_year	[[83451488.15308109, 30026648.857333377, 11448...	7477	0	float64	0.00
6	total_length_of_relationship	[[6, 3, 4, 7, 5, 2, 1, 8, 9]]	9	0	int64	0.00
5	total_days_since_first_agreement	[[2246, 1881, 1516, 2611, 2977, 3342, 3707, 21...	243	0	int64	0.00
4	total_days_since_last_agreement	[[420, 1150, 239, 116, 785, 1881, 1334, 1516, ...	208	0	int64	0.00
3	total_num_agrmts_with_parents	[[6, 15, 5, 30, 7, 10, 28, 9, 20, 11, 8, 19, 1...	37	0	int64	0.00

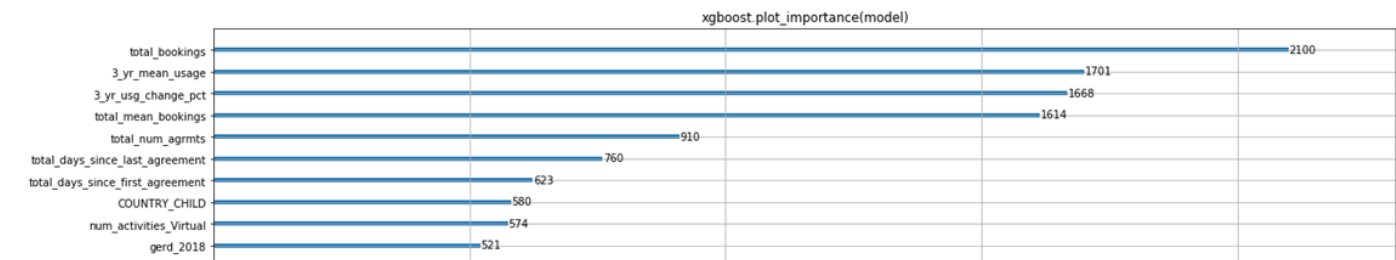
2	total_num_agrmts	[[15, 17, 10, 31, 8, 7, 35, 16, 11, 6, 9, 5, 3...	42	0	int64	0.00
18	CONSORTIUM	[[CONSORTIUM, NON-CONS]]	2	0	object	0.00
19	COUNTRY_CHILD	[[France, United States, China, Netherlands, T...	44	0	object	0.00
20	HIERARCHY_TYPE	[[ELS, RINGGOLD]]	2	0	object	0.00
21	HIER_LEVEL	[[1, 4, 3, 2, 0, 5]]	6	0	int64	0.00
36	pageviews_3_year_change	[[0.0]]	1	0	float64	0.00
35	pageviews_trend	[[no_traffic_data]]	1	0	object	0.00
34	time_3_year_change	[[0.0]]	1	0	float64	0.00
33	time_trend	[[no_traffic_data]]	1	0	object	0.00
32	visits_3_year_change	[[0.0]]	1	0	float64	0.00
31	visits_trend	[[no_traffic_data]]	1	0	object	0.00
30	3_yr_usg_change	[[no_usage_data, over_75pc_usg_dec, less_25pc_...	9	0	object	0.00
29	3_yr_usg_change_pct	[[0.0, -99.88878227367717, -99.88999404134391,...	8928	0	float64	0.00
28	3_yr_mean_usage	[[0.0, 636346.4, 7569.200000000001, 2551850.8,...	8952	0	float64	0.00
27	jnl_usage_trend	[[no_usage_data, updown_trend, increasing_tren...	4	0	object	0.00
26	CHURN_TYPE	[[NONE, PARTIAL, TOTAL]]	3	0	object	0.00
25	cust_prod_booking_trend	[[cust_prod_booking_updown_trend, cust_prod_bo...	3	0	object	0.00
24	cust_booking_trend	[[cust_booking_updown_trend, cust_booking_incr...	3	0	object	0.00
23	max_hier	[[6, 4, 5, 3, 7, 2, 1, 0]]	8	0	int64	0.00
22	num_child	[[1532, 399, 213, 353, 341, 51, 203, 14, 634, ...	122	0	int64	0.00
77	gerd_yoy_change	[[0.003567204, 0.0215608, 0.02380919, 0.087698...	120	0	float64	0.00

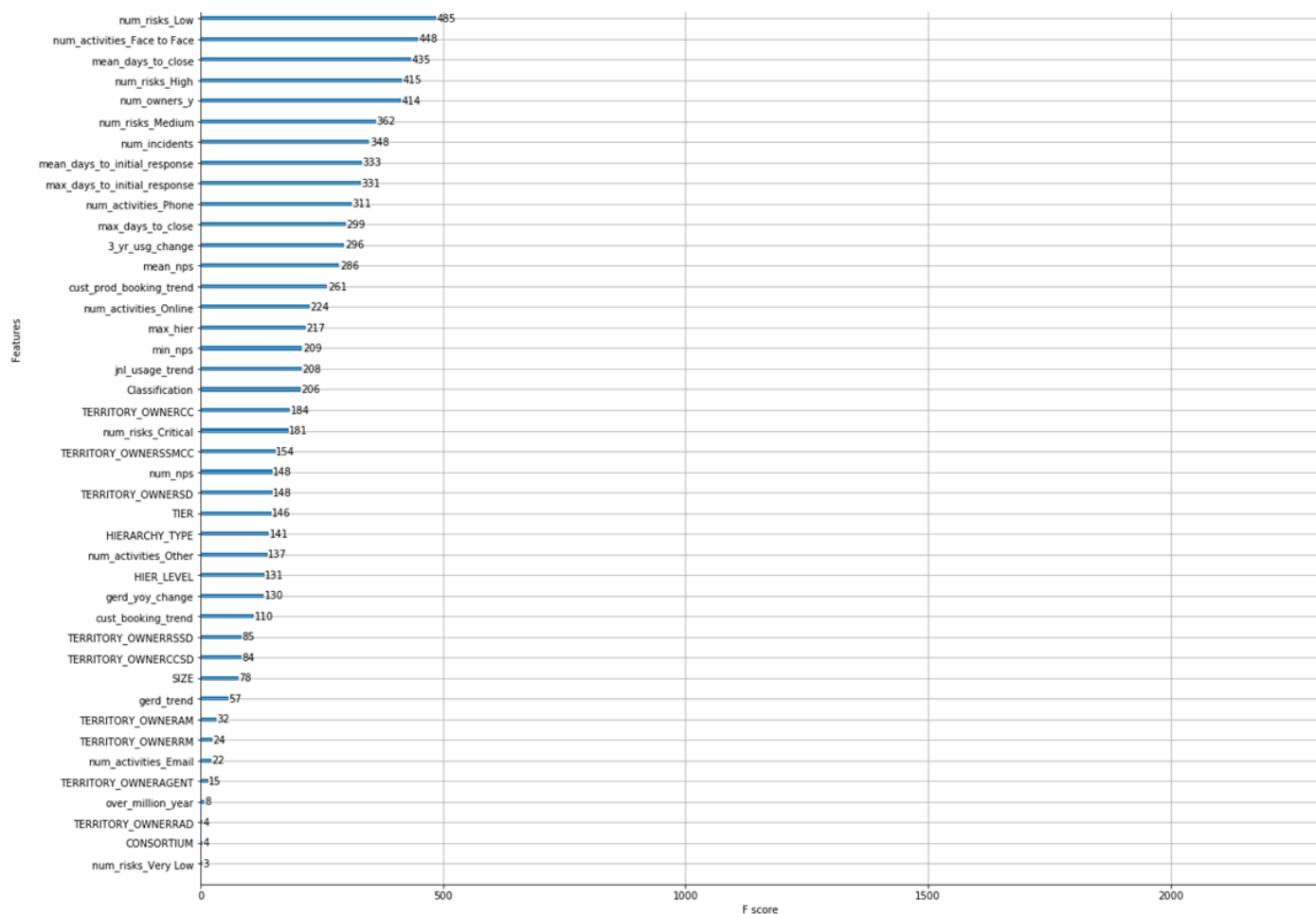
## Feature Importance:

First we develop a model taking into consideration all the engineered features, following are the feature importance we discovered.

### Feature Importance by Weights

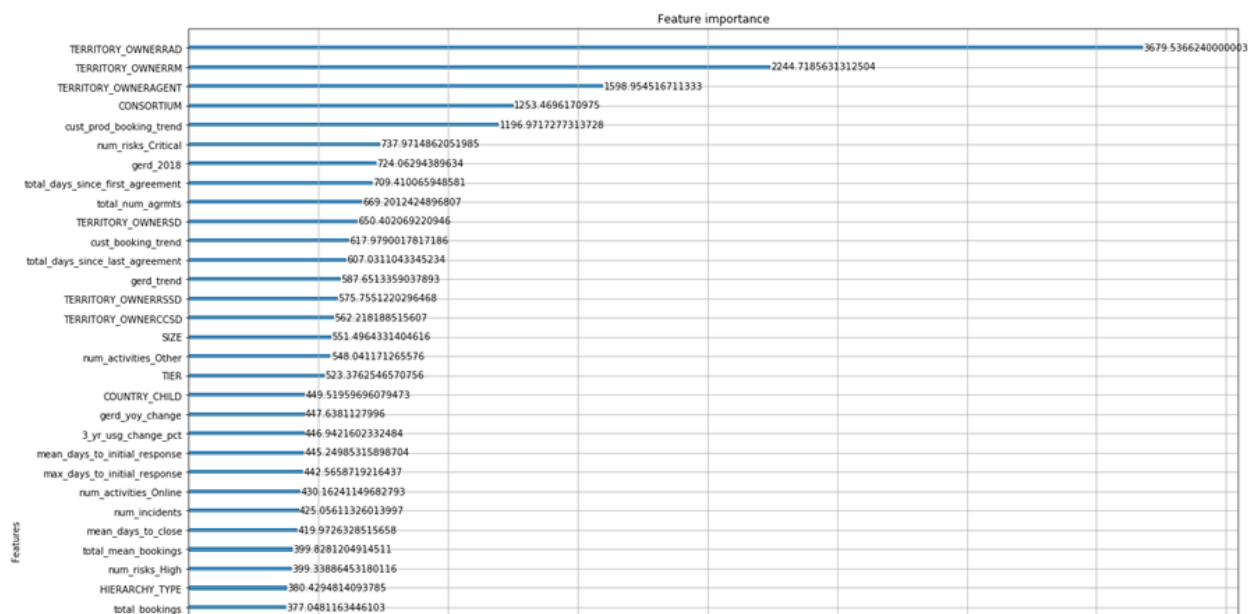
The number of times a feature is used to split the data across all trees.

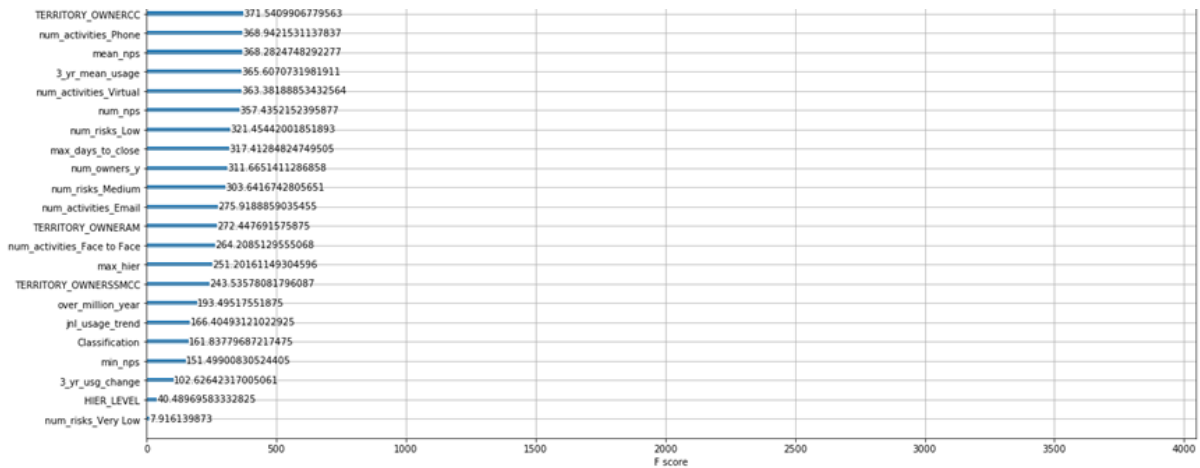




## Feature Importance by Cover

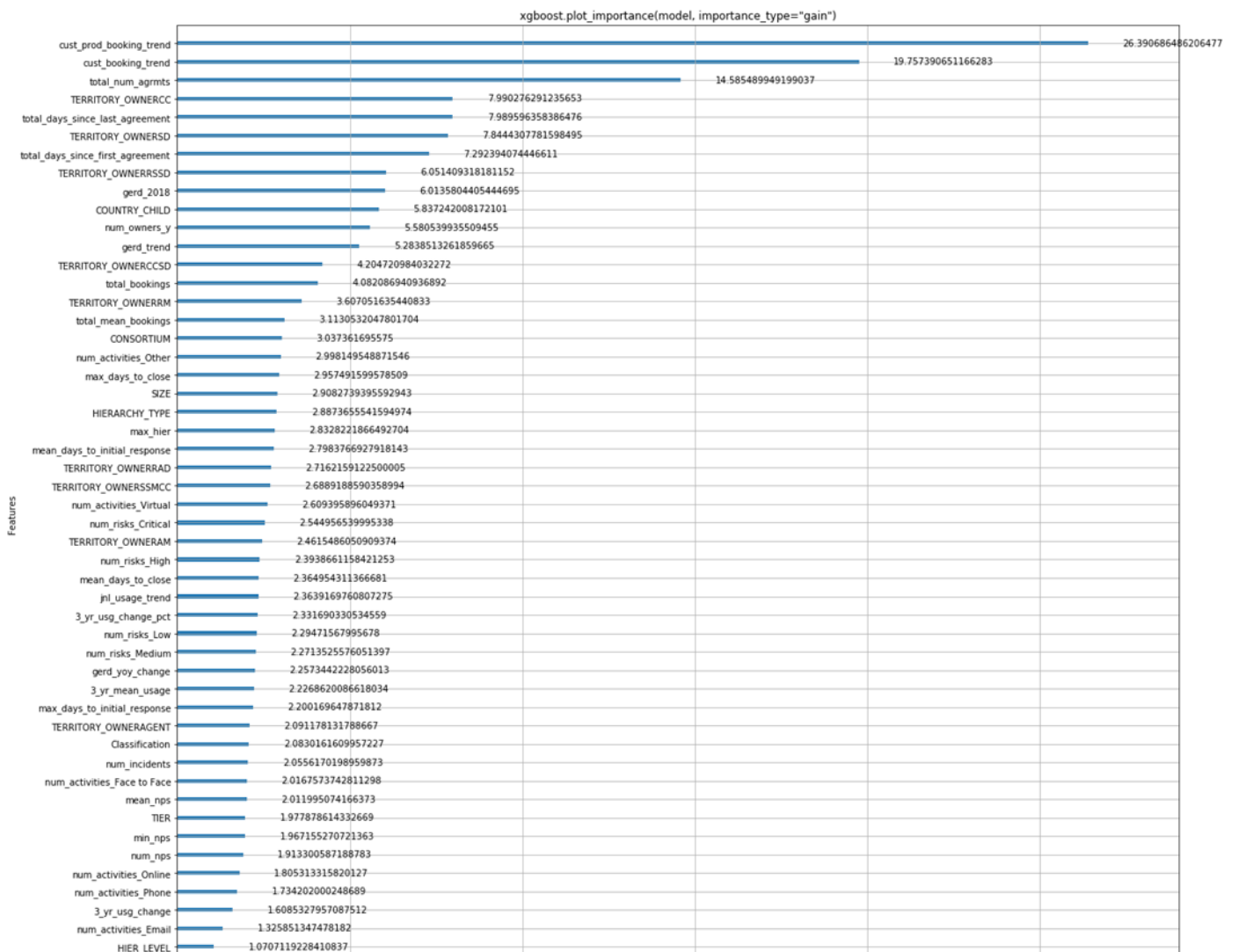
The number of times a feature is used to split the data across all trees weighted by the number of training data points that go through those splits.





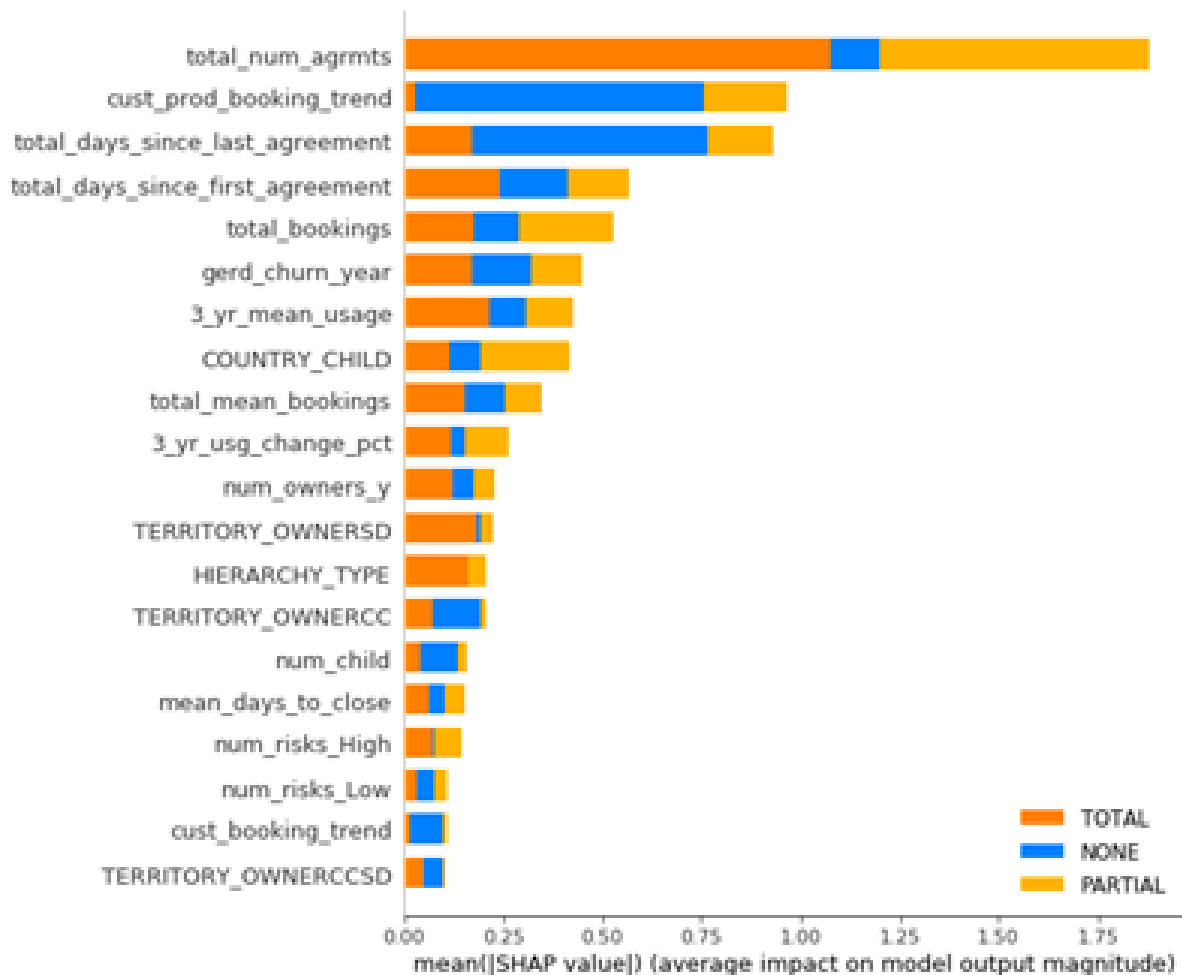
## Feature Importance by Gain

The average training loss reduction gained when using a feature for splitting.





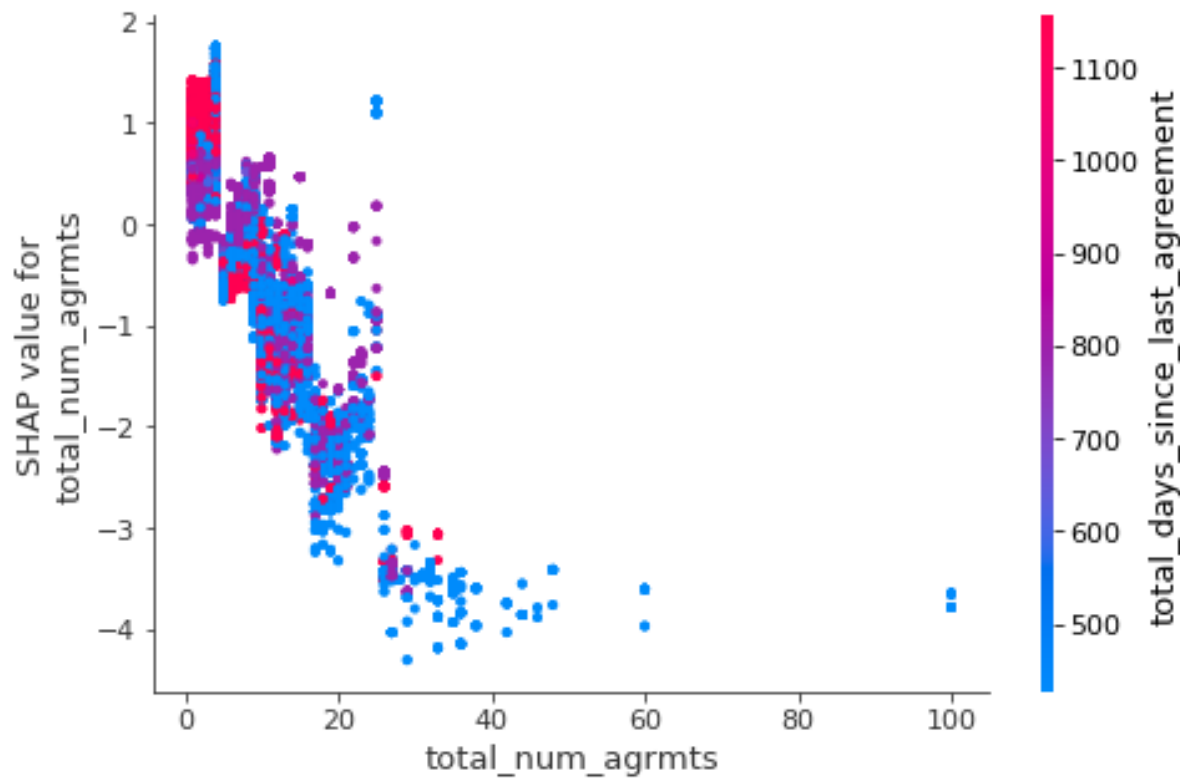
### Overall feature importance using Shapely Additive values



We see that few features have more impact than other features. To simplify the model, we keep the following 7 features:

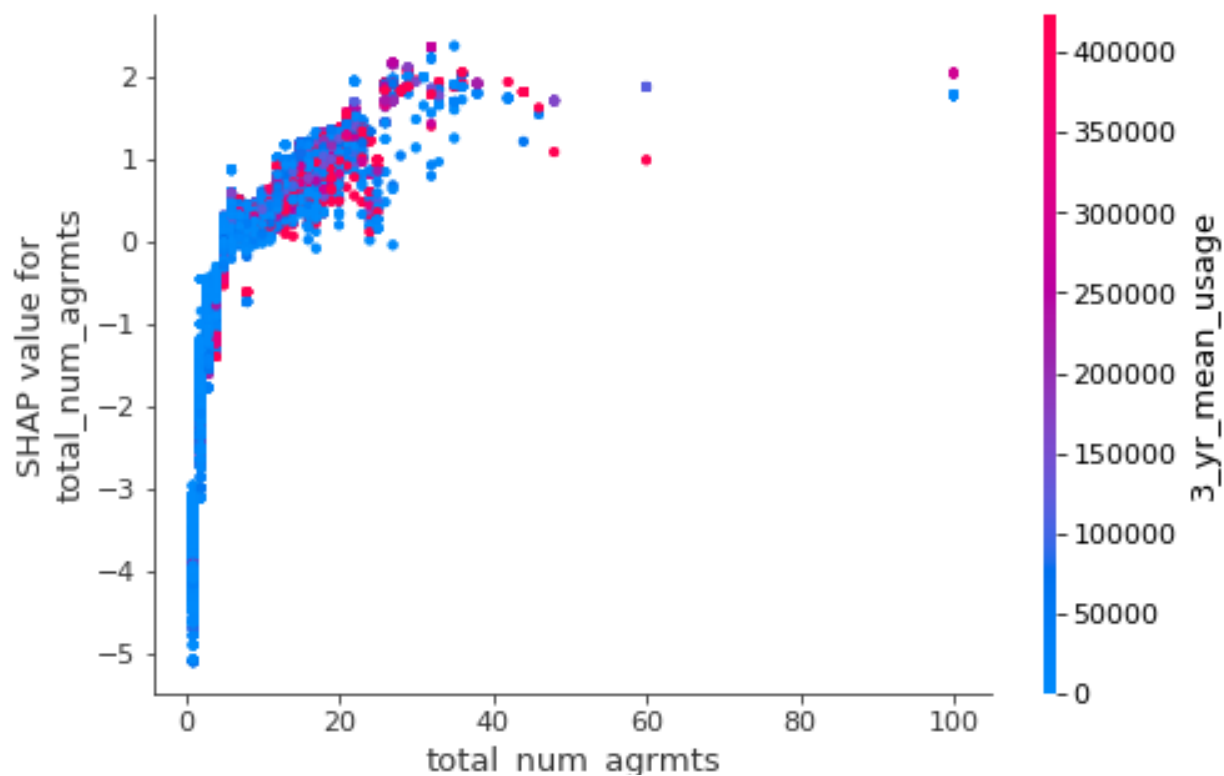
- 'total\_num\_agrmts',
- total\_days\_since\_last\_agreement (recency)
- total\_days\_since\_first\_agreement (length of relationship)
- cust\_prod\_booking\_trend
- 'total\_bookings', (size of customer spend)
- '3\_yr\_mean\_usage',
- '3\_yr\_usg\_change\_pct'
- 'gerd\_churn\_year'

it is interesting that Total Number of Agrmeents is one of the prominent predictors of churn, lets take a look at the dependence plot for total number of agreements



Dependence plot for TOTAL\_CHURN, here we see that customers with lower number of agreements are more likely to churn totally. this reduces significantly as customers have 15 agreements or more.

customers who have been with Elsevier longer and have multiple subscriptions are less likely to churn totally.





Dependence plot for total number of agreements for PARTIAL CHURN. here the story is opposite, the lower the number of agreements the less likely a customer is likely to churn partially.

however with length of relationship and more number of agreement, customers is more likely to partially churn. this could be an indicator of customers getting discounts or consolidating agreements..

## Training Method

The problem to identify customers propensity to churn is essentially a classification problem, we explored multiple classification algorithms. one of the key requirements of the model was interpret ability and simplicity

During the modelling experimentation we experimented with multiple classification models including but no limited to Gradient Boosting Classifier (GBC), Support Vector Machines (SVM), Random Forest Classifier, K Nearest Neighbor Classifier (KNN), Logistic Regression (LR).

in the initial experimentation to develop a baseline model, we observed the following accuracy levels.

Gradient Boosting Classifier:	0.82
Support vector machine(SVM):	0.53
Random Forest Classifier:	0.91
K Nearest Neighbor Classifier:	0.69
Logistic Regression:	0.53

Random Forest had the highest accuracy but this could be result of a over fitting. Where as SVM, and LR performed only slightly better than flipping a coin.

KNN model seemed to have a good result, but KNN model classifies customers with distance metric in high dimensional space and there is no feature importance.

we want a model that is relatively simple and is interpretable. one more reason for GB classifier is that these are ensembles of trees and Decision Trees do not require normalization of their inputs;

We choose XGBoost for our model development.

XGBoost stands for **eXtreme Gradient Boosting**. is an optimized distributed gradient boosting library

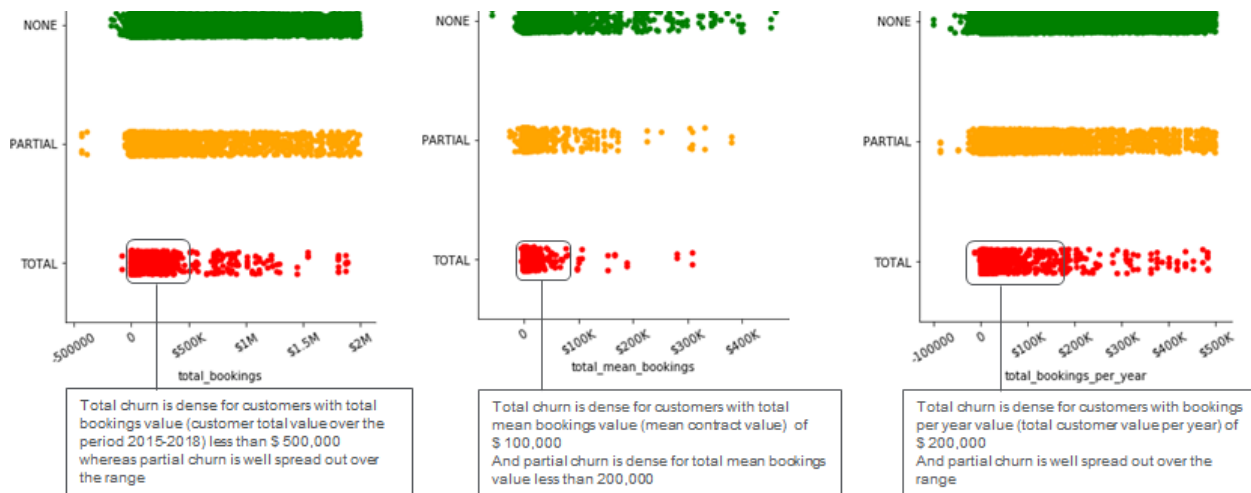
Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made. Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.

Decision Trees do not require normalization of their inputs; and since XGBoost is essentially an ensemble algorithm comprised of Decision Trees, it does not require normalization for the inputs either.

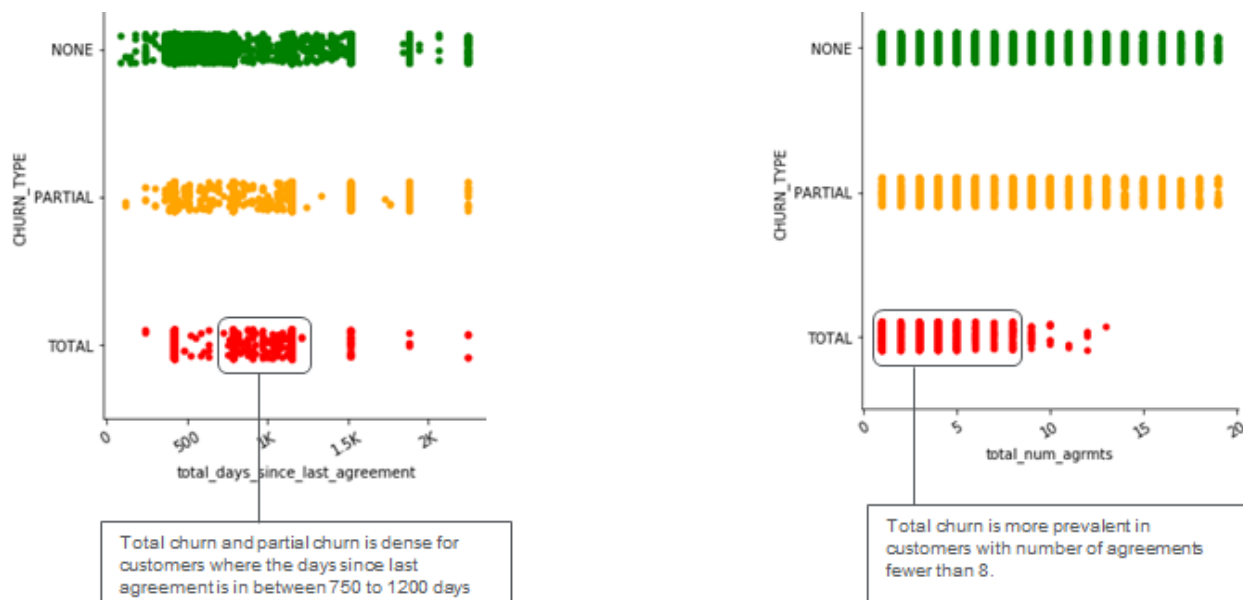
## Interesting findings

Churn by bookings value

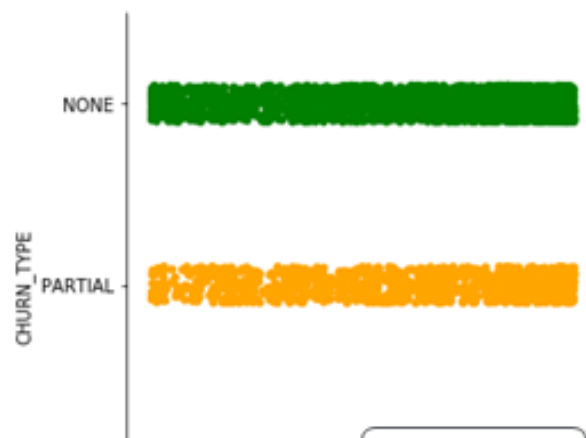


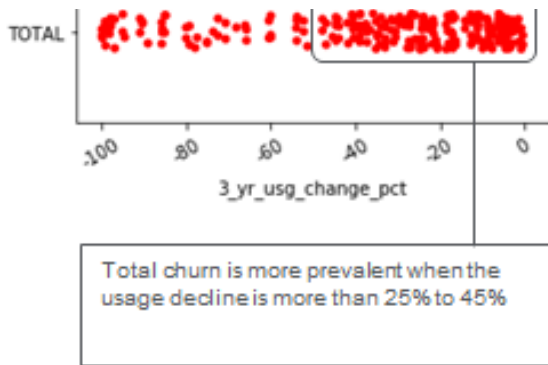


### Churn by recency and number of agreements



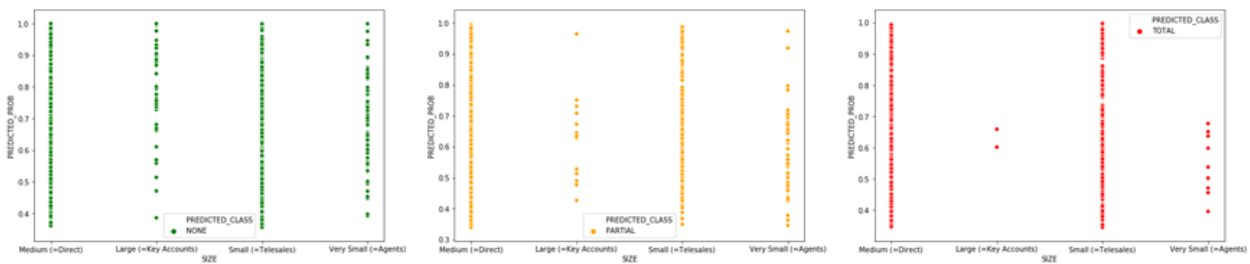
### Churn by usage



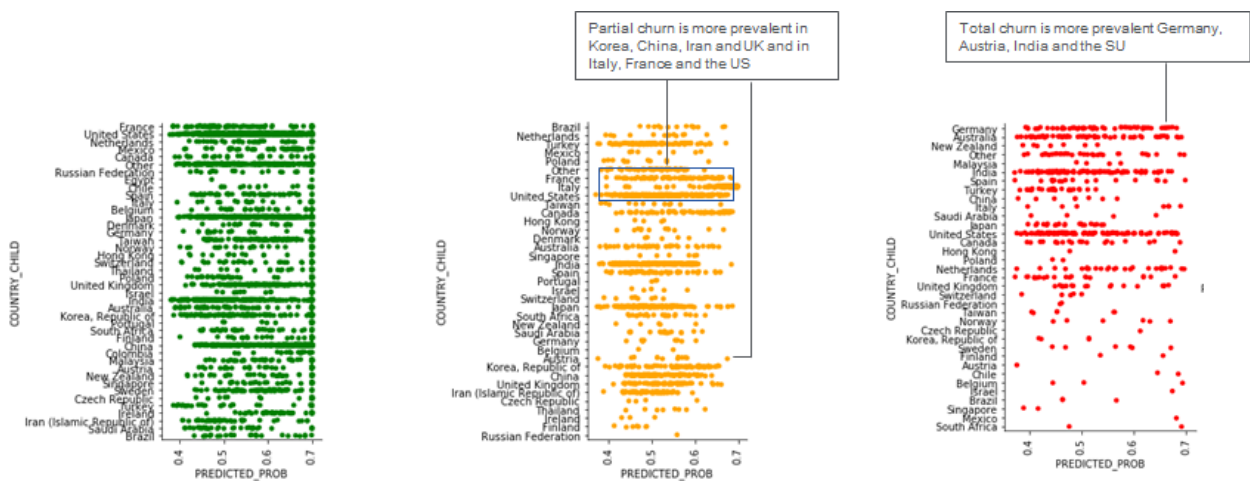


- Total churn is more prevalent when the journals usage decline is in the range 25% to 45%
- When the usage decline is more than 45%, total churn is rarefied
- This could be an indicator that when the usage decline is less than 45% customer had been using journals hut lately is not seeing value and is at high risk to churn totally
- Whereas when the usage has declined by more than 45% percent, users are not actively using the product but not concerned about cancelling the subscription. It might not be a good idea to try to engage these customers, unless there is a substantial improvement in product.

### Churn by organisation size category



### Churn by country



## **Bottlenecks**

**Availability of data:** ideally data sets would be available at the start of the project so that team can spend more time on feature engineering and model tuning and results preparation.

**Availability of tools:** Dealing with large datasets such as Usage data, we would quite often run into memory issues on the laptop, having tools such as Databrick can be leveraged to work with larger datasets.

**Deployment of model:** while the model is developed, deploying the model is a challenge given the lack of existing process to take models to production, it is imperative to establish an model productionisation solution

**Logging Accurate data:** Sales representative can be encouraged to log correct and complete data on sales force as these can provide insights and features towards model development