Great Oncology Hackathon 2020

Team 3 – Afinitor:

Rahul Agarwal
Preethika Chivukula
Srija Mukhopadhyay
Vainavi Alva
Paritosh Kulkarni

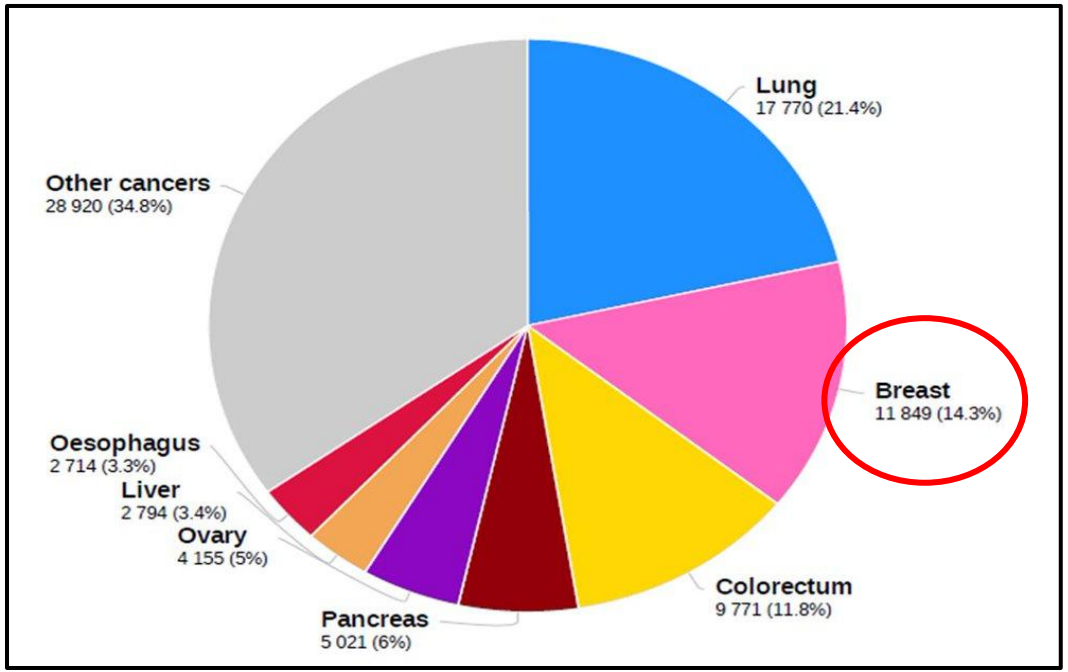
COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

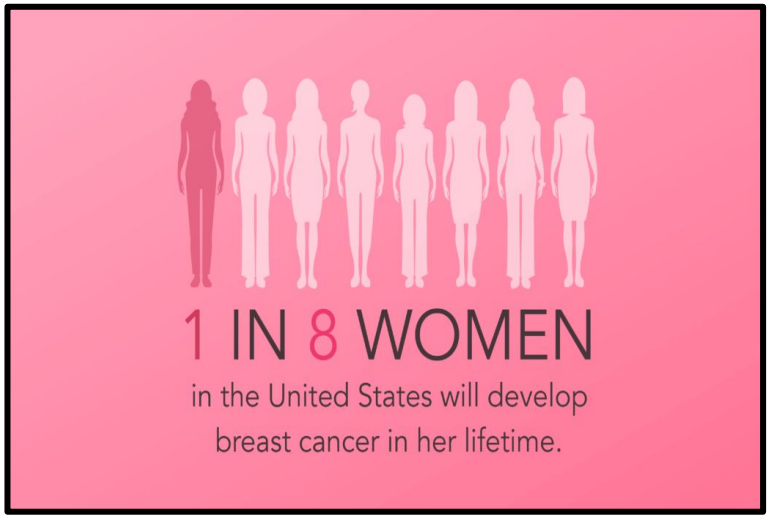
NOVARTIS | Reimagining Medicine

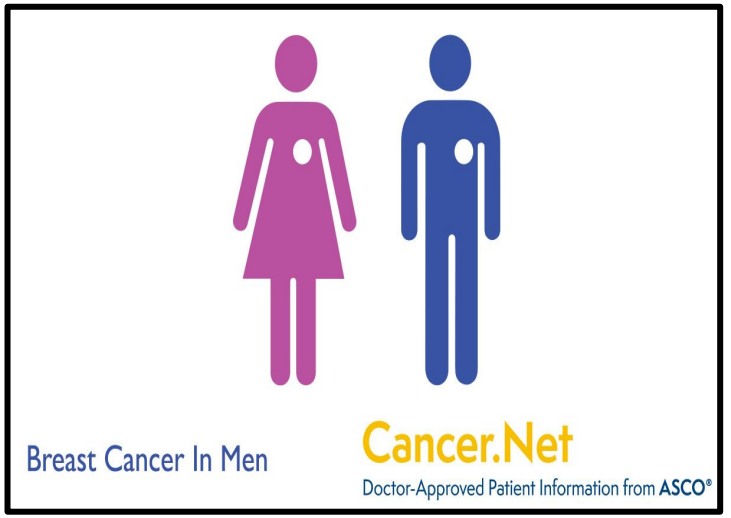
Agenda

- **Overview**
- **Problem Statement 1**
  - Challenges
  - Methodology
  - Execution Data Flow
- **Problem Statement 2**
  - Challenges
  - Methodology
  - Execution Data Flow

Facts







- About **6% of women and 9 % of men** have metastatic breast cancer when they are first diagnosed.
- Metastatic breast cancer (stage IV) is the most advanced stage of cancer
- Estimated that there are more than **168,000 women living with MBC in the US**
- **MBC easily become resistant to drug therapies.**

Problem Statement

1. *Accurately predict when a patient may be clinically diagnosed with metastatic breast cancer.*

2. *Accurately predict when a patient is likely to progress to the next "line of therapy"*

Problem Statement - 1

1. *Accurately predict when a patient may be clinically diagnosed with metastatic breast cancer.*

Problem Statement 1 - Challenges?

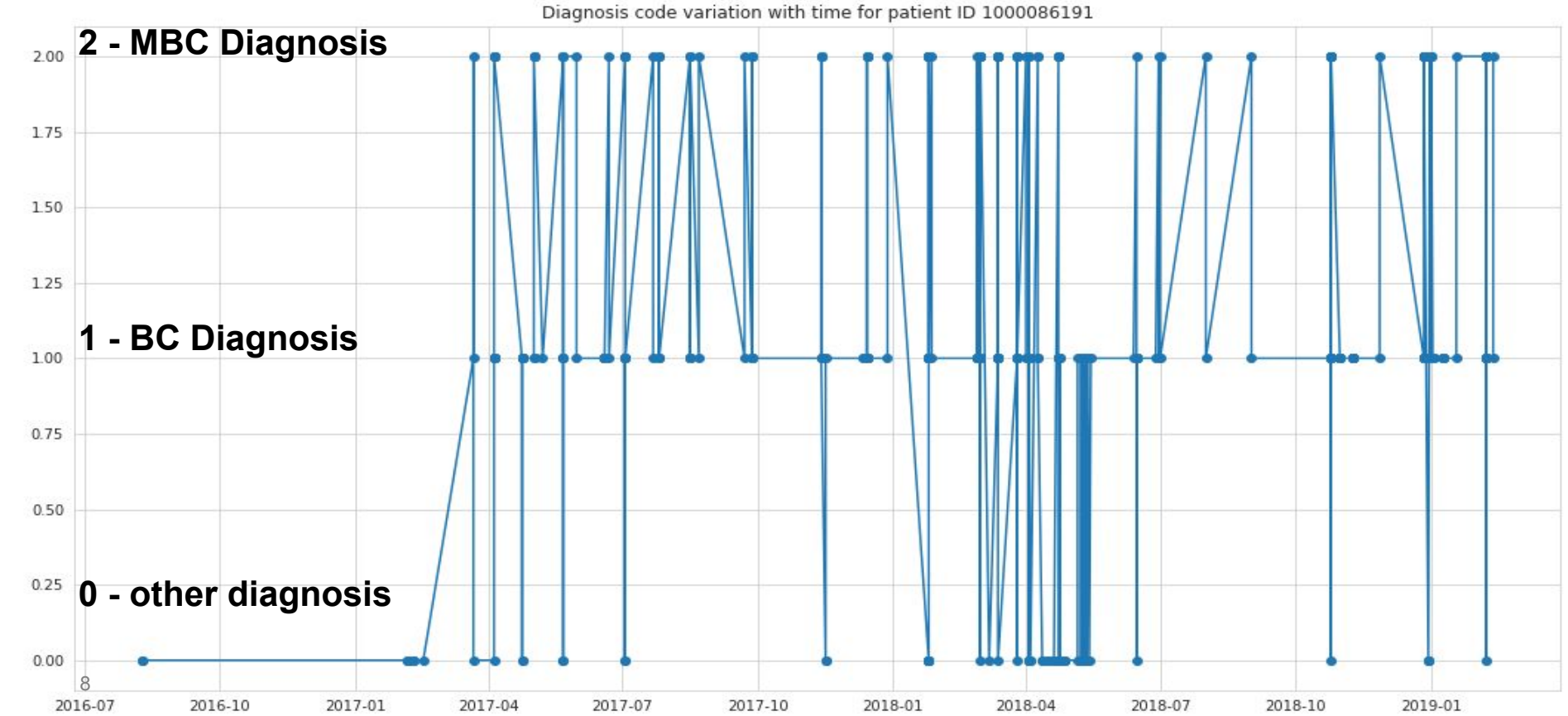
1. **Identification of MBC patients and Non-MBC patients**

   Difference between the Minimum date diagnosed with BC and minimum date when diagnosed with MBC is greater than -30

2. **Creating a holistic picture for each patient**

- Data contained multiple rows for each patient having multiple procedures, diagnosis and drugs which can be on the same date as well.
- Even the diagnosis information for any one patient would be fluctuating.

- Each patient can be diagnosed with multiple codes, multiple times throughout her life.



Methodology

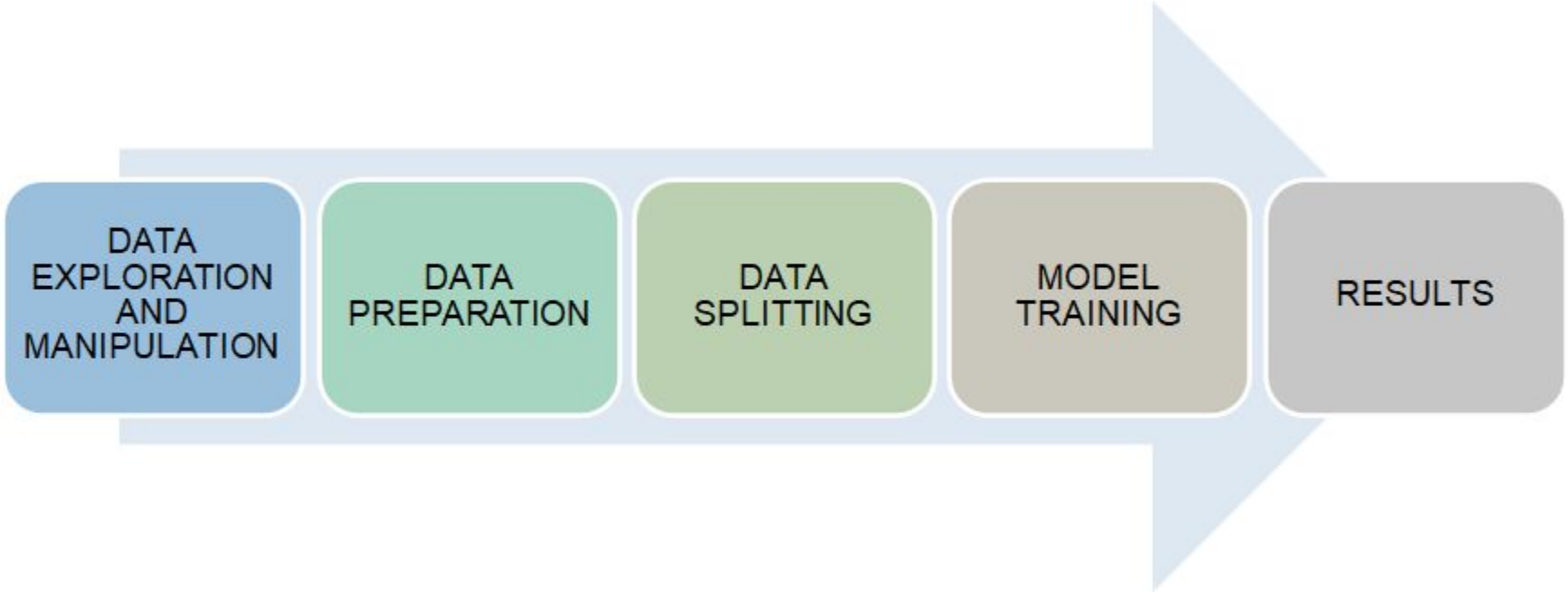
- The given data, had information on what diagnosis/procedures were already performed on the patient with their dates.
- Aimed at clubbing the time series for each patient data into one row per patient.
  - Conceretly able to identify whether the patient has MBC.
  - Create new features which extract time information etc.
- We created dummy variables for each category (like procedures, drugs etc) then did grouping based on patient, and applied summation on the dummy columns. To accumulate all the information in one cell.

Problem Statement -1 - expected outcome?

- Final output will be replicable for any new patient,
- Information required:
  - We need the history of the patient what all drugs were given, in what quantities, who is paying the bills, was there any customization done from the standard procedure, what all procedures were performed and how many times each one.

- **We feed this information to the model and the model should tell if that patient is diagnosed with MBC?**

Execution Data Flow



Data Exploration and Manipulation

- 3 tables - Dx, Px, Rx,
- Dx - no new features present compared to Rx and Px

```
dx = ['PATIENT_ID', 'CLAIM_ID', 'CLAIM_TYP_CD', 'SERVICE_DATE', 'MONTH_ID',
      'DIAGNOSIS_CODE', 'DIAG_VERS_TYP_ID', 'DIAG_CD_POSN_NBR', 'PROVIDER_ID',
      'RESTATE_FLAG', 'FLEXIBLE_FLD_1_CHAR', 'FLEXIBLE_FLD_2_CHAR']

px = ['PATIENT_ID', 'CLAIM_ID', 'CLAIM_LINE_ITEM', 'CLAIM_TYP_CD',
      'PROCEDURE_CODE', 'PRC1_MOD_CD', 'PRC1_MOD_DESC', 'PRC_VERS_TYP_ID',
      'PROVIDER_BILLING_ID', 'PROVIDER_FACILITY_ID', 'PROVIDER_REFERRING_ID',
      'PROVIDER_RENDERING_ID', 'SVC_CRGD_AMT', 'SERVICE_DATE', 'MONTH_ID',
      'UNIT_OF_SVC_AMT', 'PLACE_OF_SERVICE', 'PAYER_PLAN_ID', 'PAY_TYPE',
      'NDC', 'PRODUCT', 'DIAGNOSIS_CODE', 'DIAG_CD_POSN_NBR',
      'DIAG_VERS_TYP_ID', 'DIAG_DESC', 'WEEK_END_FRI', 'RESTATE_FLAG',
      'FLEXIBLE_FLD_1_CHAR', 'FLEXIBLE_FLD_2_CHAR']

rx = ['CLAIM_ID', 'PATIENT_ID', 'NDC', 'PROVIDER_ID', 'DIAGNOSIS_CODE',
      'DIAG_VERS_TYP_ID', 'PAYER_PLAN_ID', 'REFILL_CODE', 'DSPNSD_QTY',
      'DAYS_SUPPLY', 'SERVICE_DATE', 'MONTH_ID', 'RESTATE_FLAG',
      'FLEXIBLE_FLD_1_CHAR', 'FLEXIBLE_FLD_2_CHAR']
```

Data Exploration and Manipulation

- Combining Rx and Px records.

```
px.shape
```

```
(9297565, 16)
```

```
rx.shape
```

```
(317323, 7)
```

```
px_rx.shape
```

```
(9614888, 17)
```

- Identification of BC and MBC diagnosis codes for each patient from the reference *BC_SN ICD Code.xlsx*.
- 6618 unique procedure codes - multiple versioning systems.
  - Grouped into the respective high level version categories
  - Final category count = 29
- 13931 unique drug codes (NDC) categorized based on:
  - Brand of the drug : 10 different brands and Others.
  - Functionality of the drug : 8 drug classes and Others.

Procedure Categories

CPT I			
CPT I	AN	00000-09999	Anesthesia Services
CPT I	INT	10000-19999	Integumentary System
CPT I	MUS	20000-29999	Musculoskeletal System
CPT I	RCHL	30000-39999	Respiratory, Cardiovascular, Hemic, and Lymphatic System
CPT I	DIS	40000-49999	Digestive System
CPT I	UG	50000-59999	Urinary, Male Genital, Female Genital, Maternity Care, and Delivery System
CPT I	EN	60000-69999	Endocrine, Nervous, Eye and Ocular Adnexa, Auditory System
CPT I	RAD	70000-79999	Radiology Services
CPT I	PATH	80000-89999	Pathology and Laboratory Services
CPT I	EMS	90000-99999	Evaluation & Management Services
CPT I	AMAAA	0002M- 0014M	Administrative Multianalyte assay (MAAA)

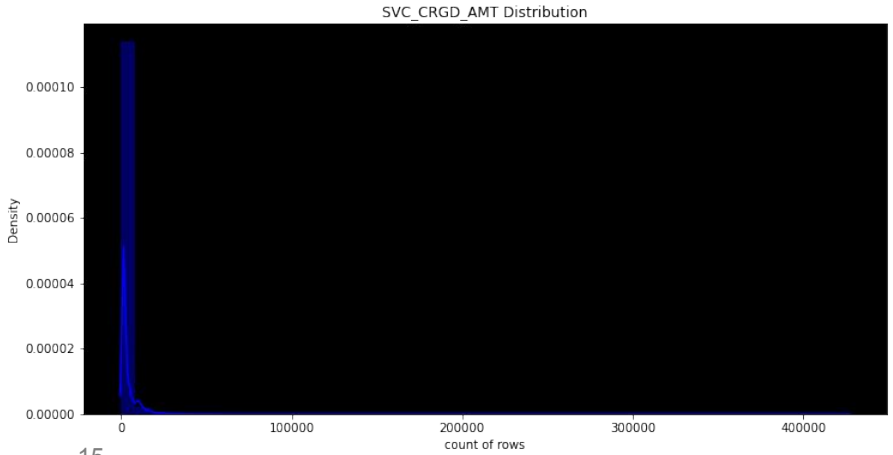
CPT II			
CPT II	CM	0001F-0015F	Composite Measures
CPT II	PM	0500F-0584F	Patient Management
CPT II	PH	1000F-1505F	Patient History
CPT II	PE	2000F-2060F	Physical Examination
CPT II	DIAG	3006F-3776F	Diagnostic/Screening Processes or Results
CPT II	THE	4000F-4563F	Therapeutic, Preventive or Other
CPT II	FOP	5005F-5250F	Follow-up or Other Outcomes
CPT II	PS	6005F-6150F	Patient Safety
CPT II	SM	7010F-7025F	Structural Measures
CPT II	NON	9001F-9007F	Non-Measure Claims Based Reporting

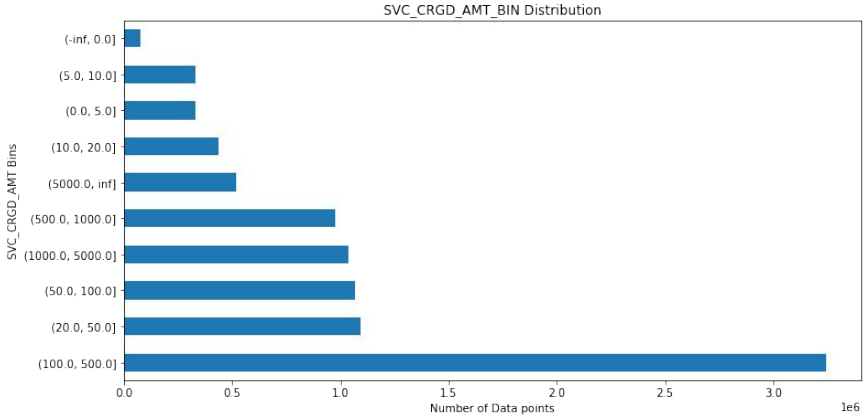
HCPCS II			
HCPCS II	II-A	A0000-A9999	Transportation, Medical & Surgical
HCPCS II	II-B	B0000-B9999	Enteral and Parenteral Therapy
HCPCS II	II-C	C0000-C9999	Temporary Hospital Outpatient Pros
HCPCS II	II-D	D0000-D9999	Dental Procedures
HCPCS II	II-E	E0000-E9999	Durable Medical Equipment
HCPCS II	II-G	G0000-G9999	Temporary Procedures & Professiona
HCPCS II	II-H	H0000-H9999	Rehabilitative Services
HCPCS II	II-J	J0000-J9999	Drugs Administered Other Than Oral
HCPCS II	II-K	K0000-K9999	Temporary Codes for Durable Medica
HCPCS II	II-L	L0000-L9999	Orthotic/Prosthetic Procedures
HCPCS II	II-M	M0000-M9999	Medical Services
HCPCS II	II-P	P0000-P9999	Pathology and Laboratory
HCPCS II	II-Q	Q0000-Q9999	Temporary Codes
HCPCS II	II-R	R0000-R9999	Diagnostic Radiology Services
HCPCS II	II-S	S0000-S9999	Private Payer Codes
HCPCS II	II-T	T0000-T9999	State Medicaid Agency Codes
HCPCS II	II-V	V0000-V9999	Vision/Hearing Services

CPT III			
CPT III	TEMP	0001T - 0999T	temporary codes that describe emerging and experimental technologies, services and procedure

Data Exploration and Manipulation

- **Drug Dosage multiplier** and the **charges related to procedures** was unevenly distributed - increased counts towards the two extremes.
  - Negative values -> absolute values
  - Performed binning : **[-np.inf, 0, 5, 10, 20, 50, 100, 500, 1000, 5000, np.inf]**





NOVARTIS | Reimagining Medicine

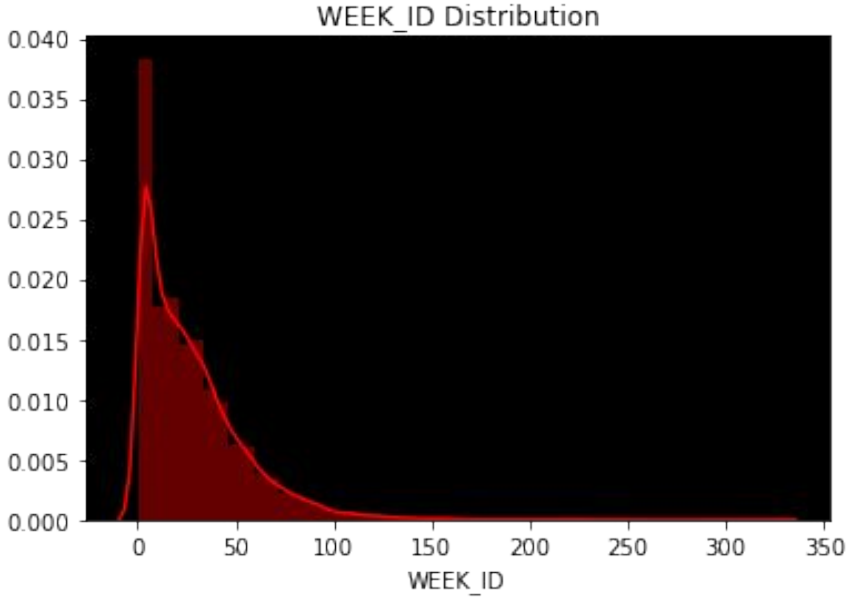
Data Exploration and Manipulation

- Some procedures -> customized modifier.
- To emphasize this aspect we created a separate binary variable:

  - **BOOL_FLEXIBLE_FLD**

  - **BOOL_PROC_MOD**

- Tracking each patient history till cancer diagnosis -
  - **WEEK_ID** -> ie the weekNumber_Year
  - Helps track the number of unique weeks the patient was under observation

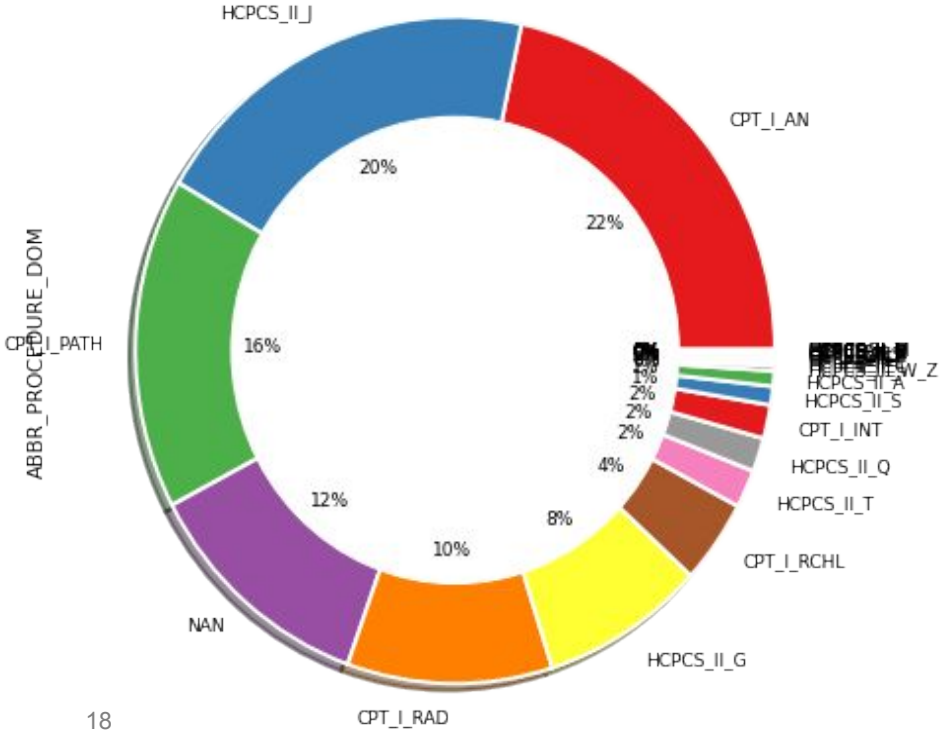


Data Preparation

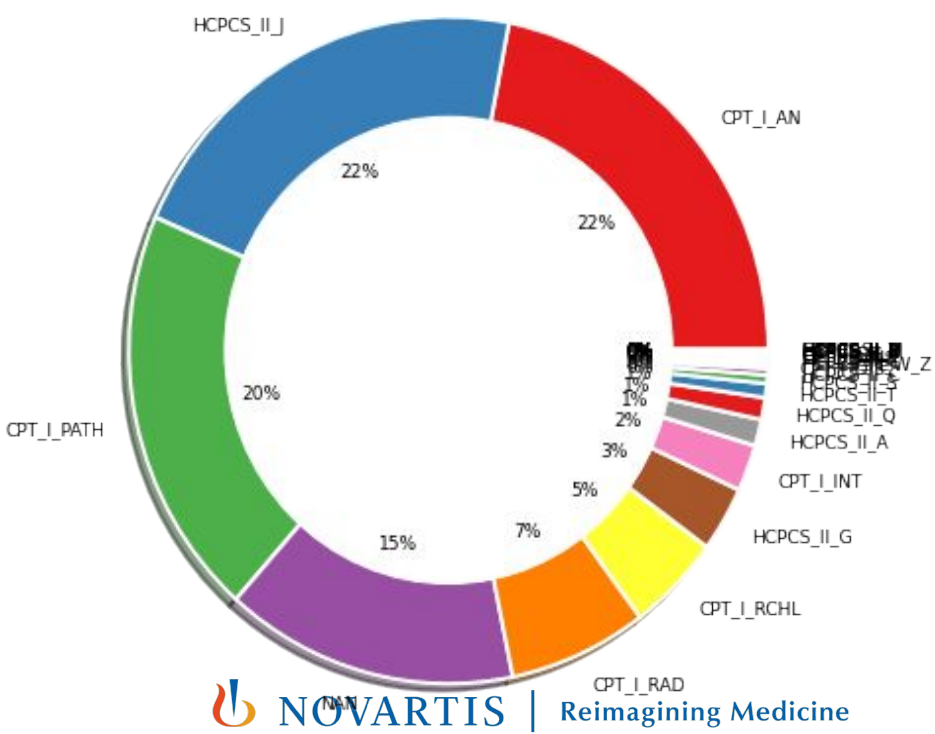
- Removed the factors which did not add any domain value to identify help in predicting the diagnosis
  - All CLAIM ID related columns
  - PATIENT_ID identifier

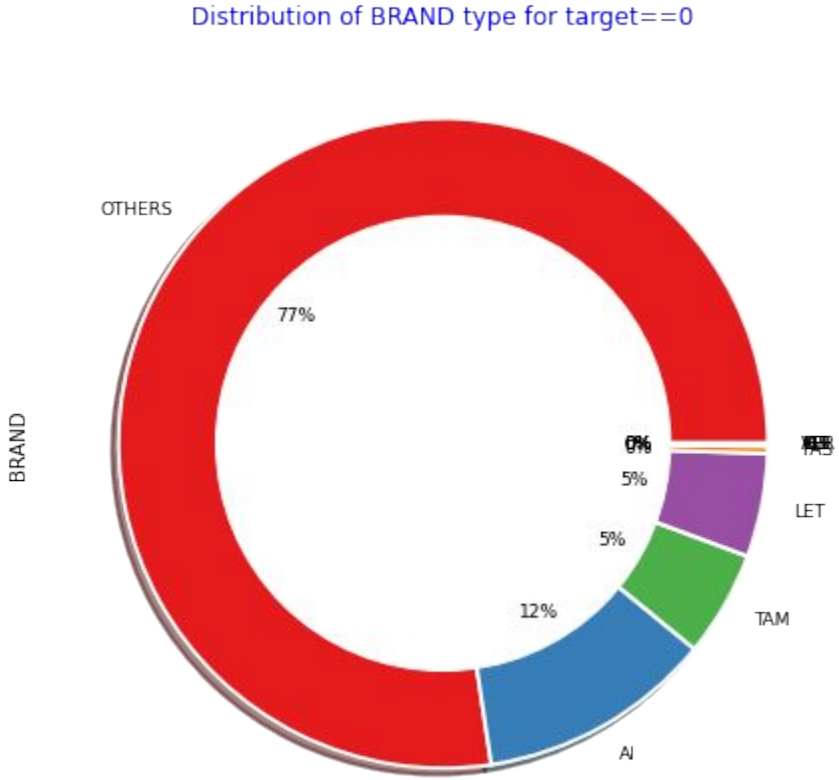
EDA

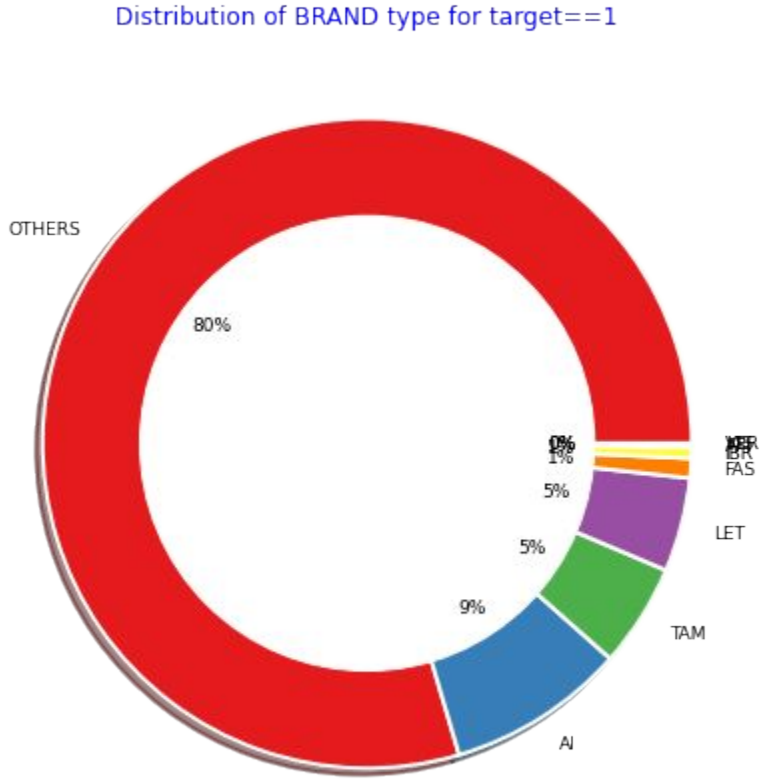
Distribution of ABBR_PROCEDURE_DOM type for target==0



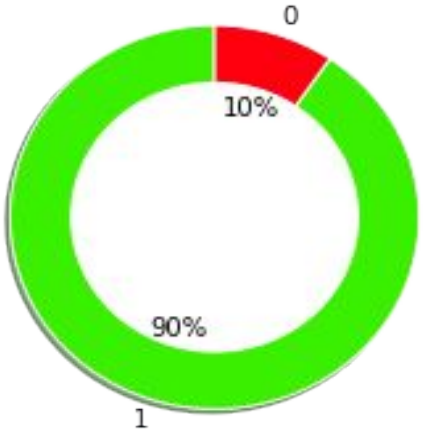
Distribution of ABBR_PROCEDURE_DOM type for target==1



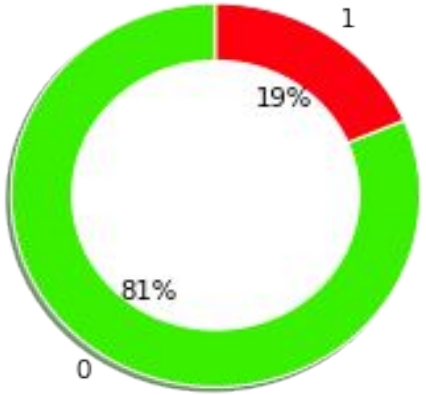




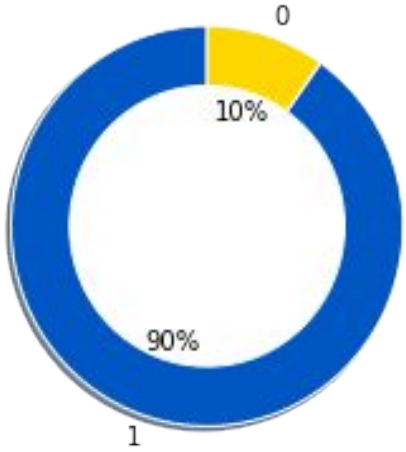
BOOL_FLEXIBLE_FLD- MBC Patient



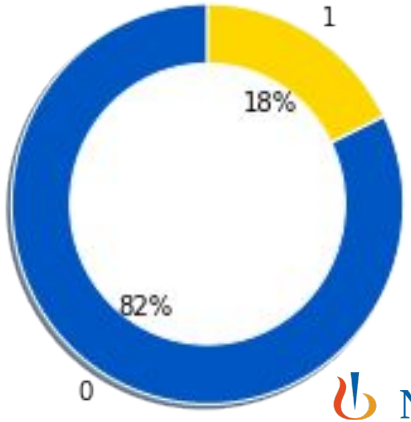
BOOL_PROC_MOD- MBC Patient



BOOL_FLEXIBLE_FLD- Non-MBC Patient



BOOL_PROC_MOD- Non-MBC Patient



Data Splitting

- For each patient - post identification of the first date of diagnosis as MBC_patient, all other data points were filtered out - avoid **information Leak.**
  - **only records with Service_date<=minimum MBC detected dates**
- Model training - split the data - Training : 75%, Testing : 25%
-

MBC	6478
Non MBC (BC or others)	8522
Total	**15000**

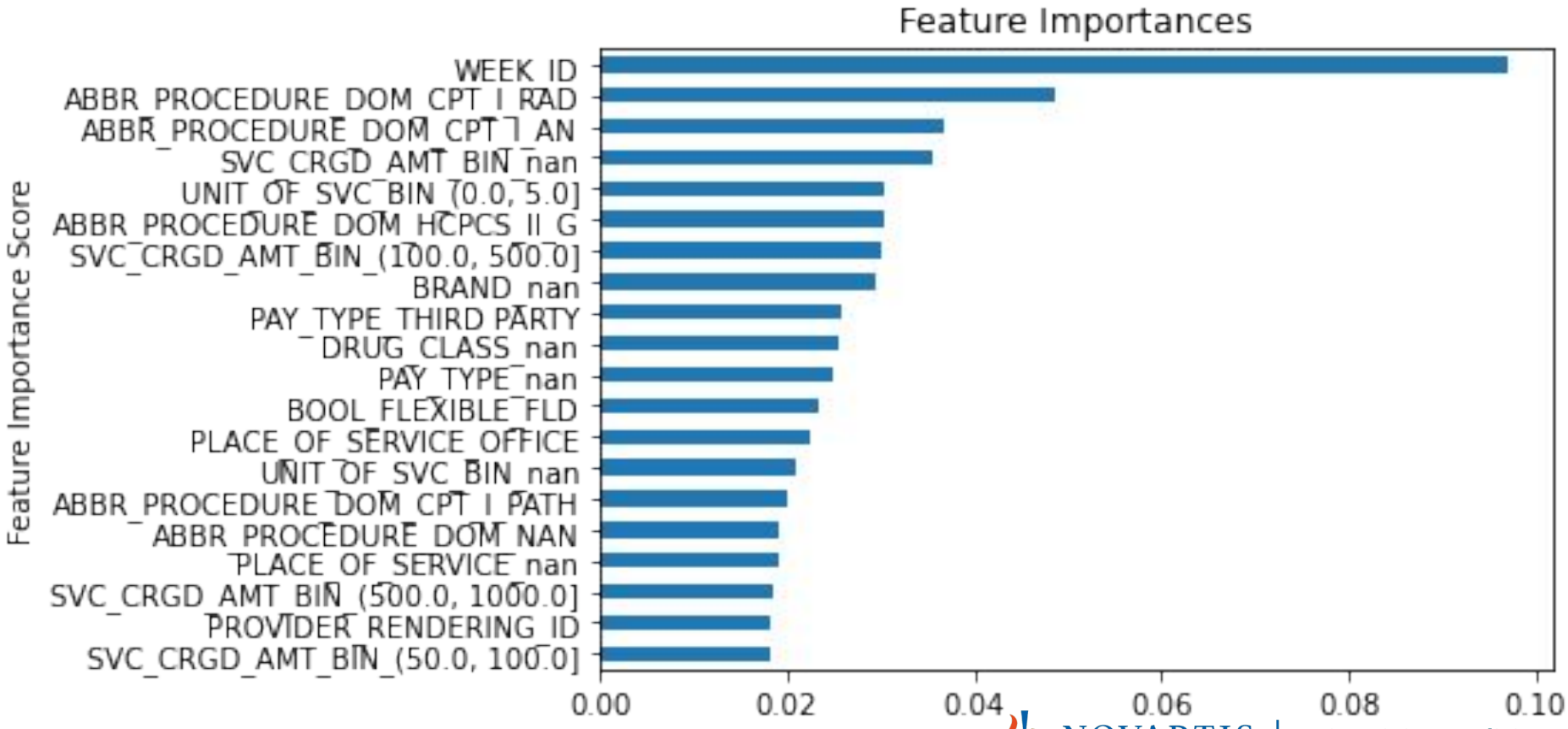
- Model training - Resampling training data to even out the target category distribution

Model Training

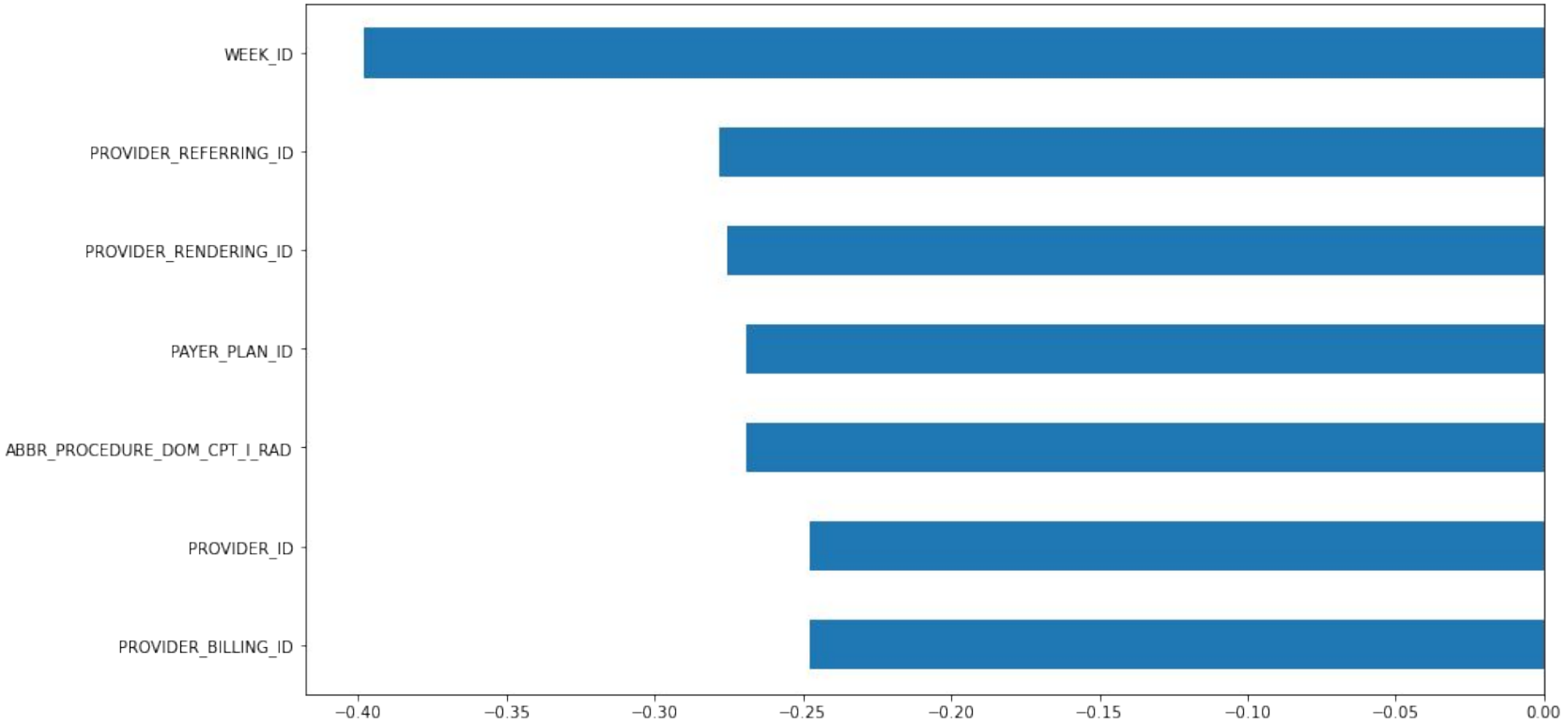
	Model	Fitting time	Precision	Accuracy	Recall/Sensitivity	Specificity	F1_score	AUC_ROC
4	Random Forest	3.10527	0.869889	0.869671	0.869742	0.899446	0.869658	0.933808
7	CatBoost	3.0612s	0.921166	0.847454	0.799005	0.776292	0.855748	0.854860
1	Decision Tree	0.454832	0.806486	0.803958	0.803690	0.784594	0.803507	0.803681
0	Logistic Regression	0.224085	0.763971	0.762810	0.762615	0.732472	0.762498	0.807196
5	K-Nearest Neighbors	0.164913	0.755307	0.755143	0.755062	0.730627	0.755084	0.828729
2	Linear Discriminant Analysis	0.198388	0.751226	0.745286	0.744832	0.665129	0.743649	0.823327
6	Bayes	0.0281436	0.641977	0.582178	0.580246	0.256919	0.530881	0.642651
3	Quadratic Discriminant Analysis	0.0710277	0.642867	0.544002	0.541536	0.198339	0.444619	0.719566

- 8 different models
- 25% test data

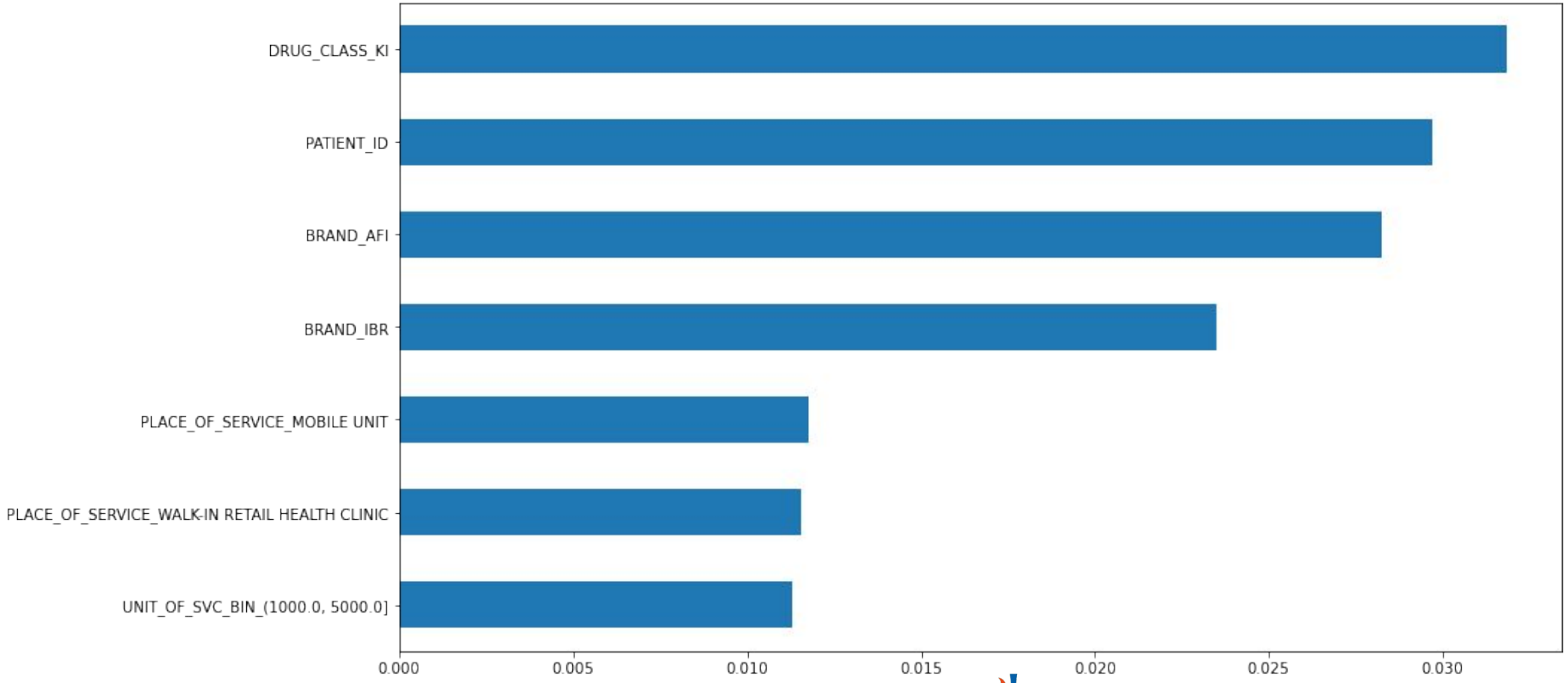
Top features from the Random Forest model



Top negatively correlated features



Top positively correlated features

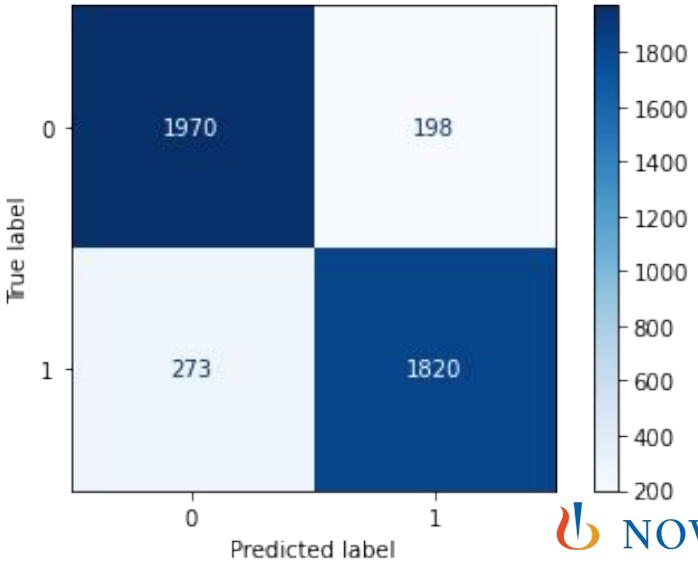


Results & Findings for Prob 1

- **Random Forest** -> most consistent results.
- **Training** - 5 fold cross validation
- **Top feature** importances used by the model ->top 10 correlated columns.

Classification report.

- **Accuracy**: *0.8696*
- **Sensitivity**: *0.8697*
- **Specificity**: *0.8999*



Problem Statement - 2

Accurately predict when a patient is likely to progress to the next "line of therapy"

Problem Statement 2 - Challenges?

1. **Change of line of treatment**

   When a patient shift's current treatment, assumption, the class of drugs given to the patient corresponds to a kind of treatment.

   Difference between the Minimum date of **first treatment and minimum date of second treatment should be greater than or equal to 30 days for change.**

2. **Creating a holistic picture for each patient**

- Data **contained multiple rows for each patient having multiple procedures**, diagnosis and drugs which can be on the same date as well.
- The **drug class for any one patient can change with time**.

Methodology

- The given data, had information on what **diagnosis/procedures were already performed** on the patient with their dates.

- Aimed at clubbing the **time series for each patient data** into one row per patient.
  - Concretely **able to identify** what treatment the patient was going through..
  - Create **new features** which extract time information etc.

- We created **dummy variables** for each category (like procedures, drugs brands etc) then did **grouping based on patient**, and applied summation on the dummy columns. To accumulate all the information in one cell.

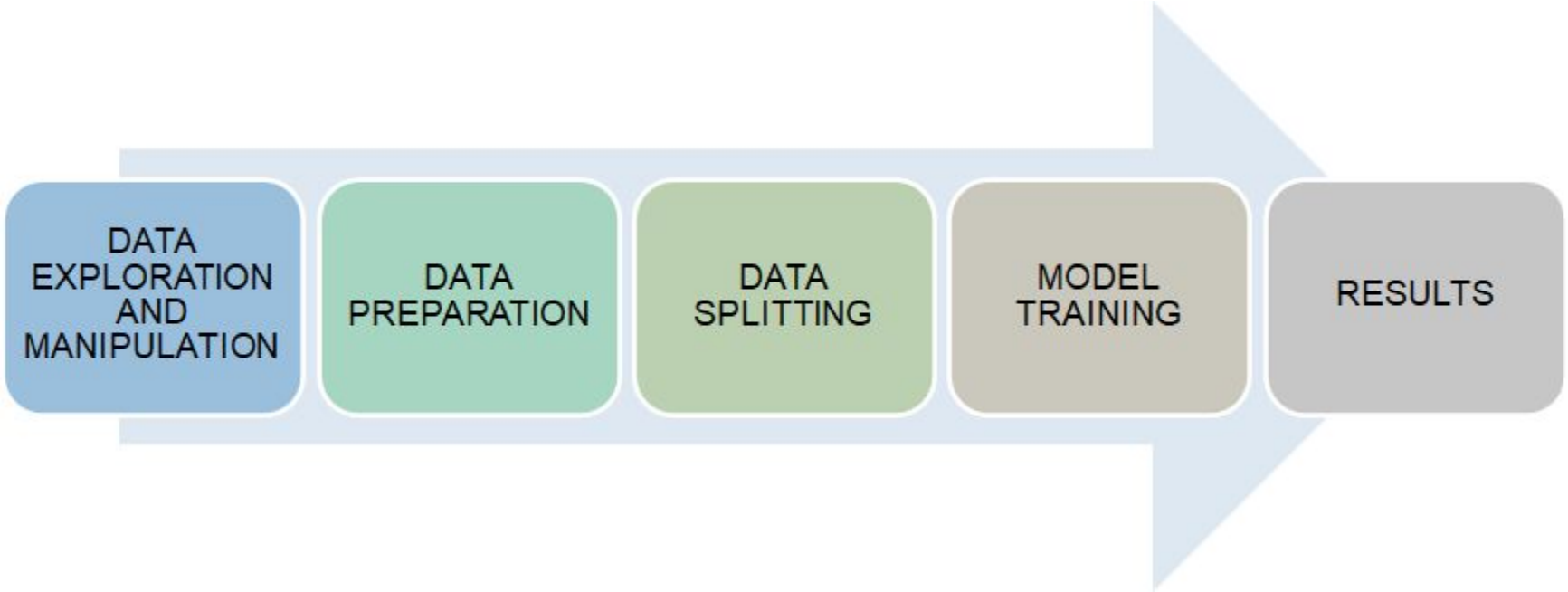
Problem Statement -2 - expected outcome?

Final output will be replicable for any new patient,

- Information required:
  - We need the history of the patient what all drugs were given, in what quantities, who is paying the bills, was there any customization done from the standard procedure, what all procedures were performed and how many times each one.

- **We feed this information to the model and the model should tell if that patient will change their line of treatment?**

Data Flow



Data Exploration and Manipulation

- 3 tables - Dx, Px, Rx,
- Dx - no new features present compared to Rx and Px

```
dx = ['PATIENT_ID', 'CLAIM_ID', 'CLAIM_TYP_CD', 'SERVICE_DATE', 'MONTH_ID',
      'DIAGNOSIS_CODE', 'DIAG_VERS_TYP_ID', 'DIAG_CD_POSN_NBR', 'PROVIDER_ID',
      'RESTATE_FLAG', 'FLEXIBLE_FLD_1_CHAR', 'FLEXIBLE_FLD_2_CHAR']

px = ['PATIENT_ID', 'CLAIM_ID', 'CLAIM_LINE_ITEM', 'CLAIM_TYP_CD',
      'PROCEDURE_CODE', 'PRC1_MOD_CD', 'PRC1_MOD_DESC', 'PRC_VERS_TYP_ID',
      'PROVIDER_BILLING_ID', 'PROVIDER_FACILITY_ID', 'PROVIDER_REFERRING_ID',
      'PROVIDER_RENDERING_ID', 'SVC_CRGD_AMT', 'SERVICE_DATE', 'MONTH_ID',
      'UNIT_OF_SVC_AMT', 'PLACE_OF_SERVICE', 'PAYER_PLAN_ID', 'PAY_TYPE',
      'NDC', 'PRODUCT', 'DIAGNOSIS_CODE', 'DIAG_CD_POSN_NBR',
      'DIAG_VERS_TYP_ID', 'DIAG_DESC', 'WEEK_END_FRI', 'RESTATE_FLAG',
      'FLEXIBLE_FLD_1_CHAR', 'FLEXIBLE_FLD_2_CHAR']

rx = ['CLAIM_ID', 'PATIENT_ID', 'NDC', 'PROVIDER_ID', 'DIAGNOSIS_CODE',
      'DIAG_VERS_TYP_ID', 'PAYER_PLAN_ID', 'REFILL_CODE', 'DSPNSD_QTY',
      'DAYS_SUPPLY', 'SERVICE_DATE', 'MONTH_ID', 'RESTATE_FLAG',
      'FLEXIBLE_FLD_1_CHAR', 'FLEXIBLE_FLD_2_CHAR']
```

Data Exploration and Manipulation

- Combining Rx and Px records.

```
px_21.shape
```

```
(11690977, 16)
```

```
rx_21.shape
```

```
(583878, 7)
```

```
px_rx_21.shape
```

```
(12274855, 17)
```

- Identification of BC and MBC diagnosis codes for each patient from the reference.

- 6325 unique procedure codes - multiple versioning systems.
  - Further grouping into the respective high level version categories
  - Final category count = 29
- 11171 unique drug codes (NDC) categorized based on:
  - Brand of the drug : 10 different brands and Others.
  - Functionality of the drug : 7 drug classes and Others.

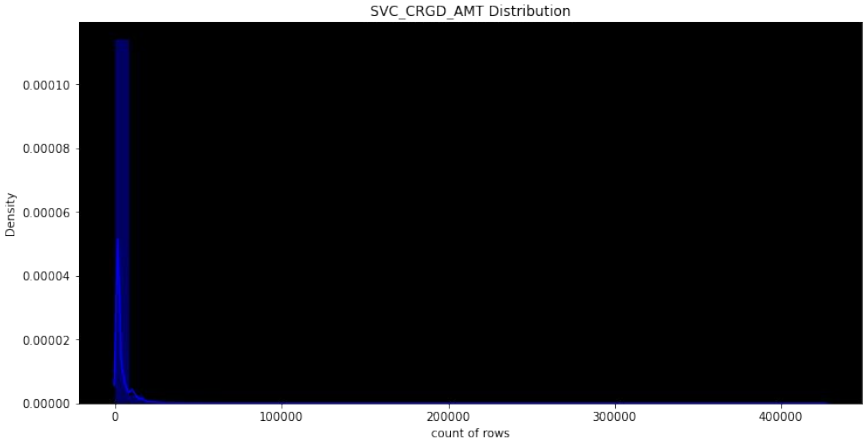
Drug Classes

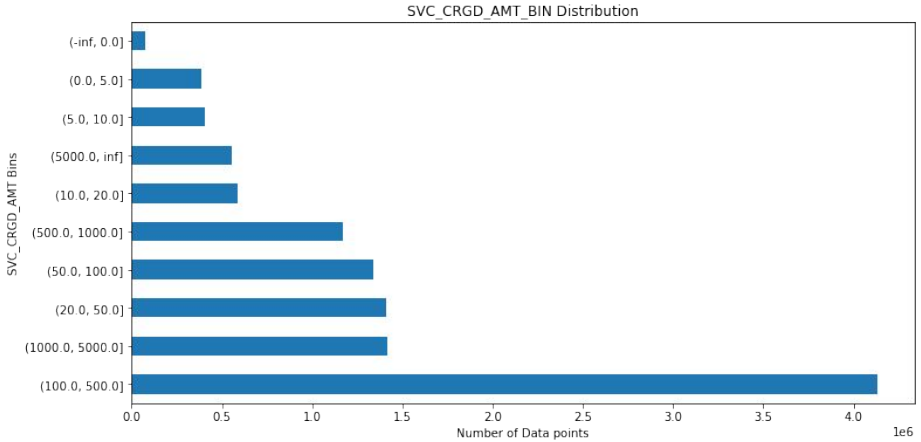
Application/Purpose based categories

KEY	
KI	Kinase inhibitor
SERM	Selective estrogen receptor modulator
SERD	selective estrogen receptor degrader
AI	Aromatase inhibitor
CD	combination drug
AM	antimetabolite
MIS	miscellaneous chemotherapy drugs
RET	retinoid

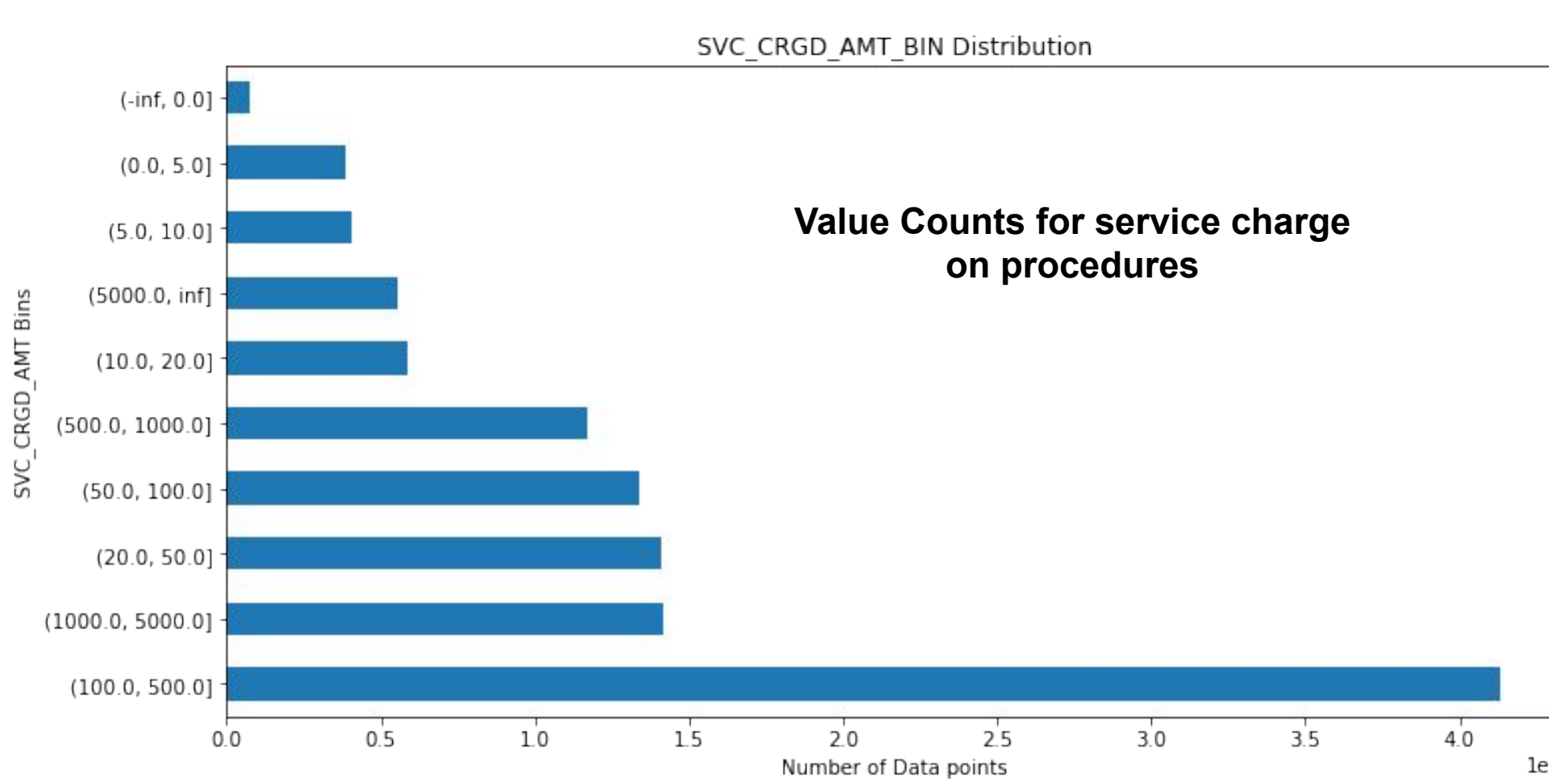
Data Exploration and Manipulation

- Drug Dosage multiplier and the charges related to procedures was unevenly distributed - increased counts towards the two extremes.
  - Negative values -> absolute values
  - Performed binning : **[-np.inf, 0, 5, 10, 20, 50, 100, 500, 1000, 5000, np.inf]**





NOVARTIS | Reimagining Medicine



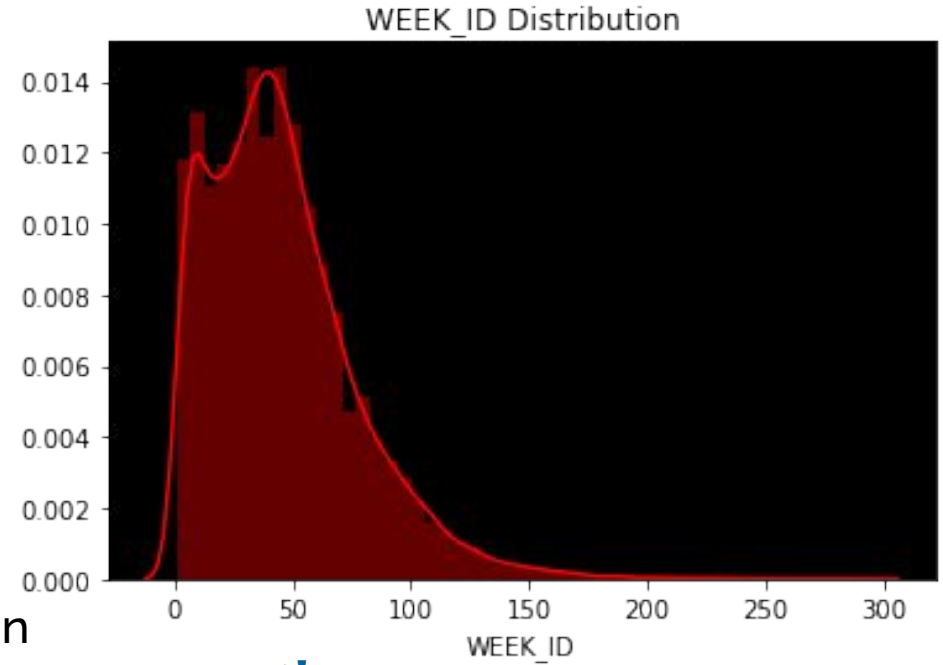
Value Counts for service charge on procedures

Data Exploration and Manipulation

- Some procedures -> customized modifier.
- To emphasize this aspect we created a separate binary variable:

  - **BOOL_FLEXIBLE_FLD**

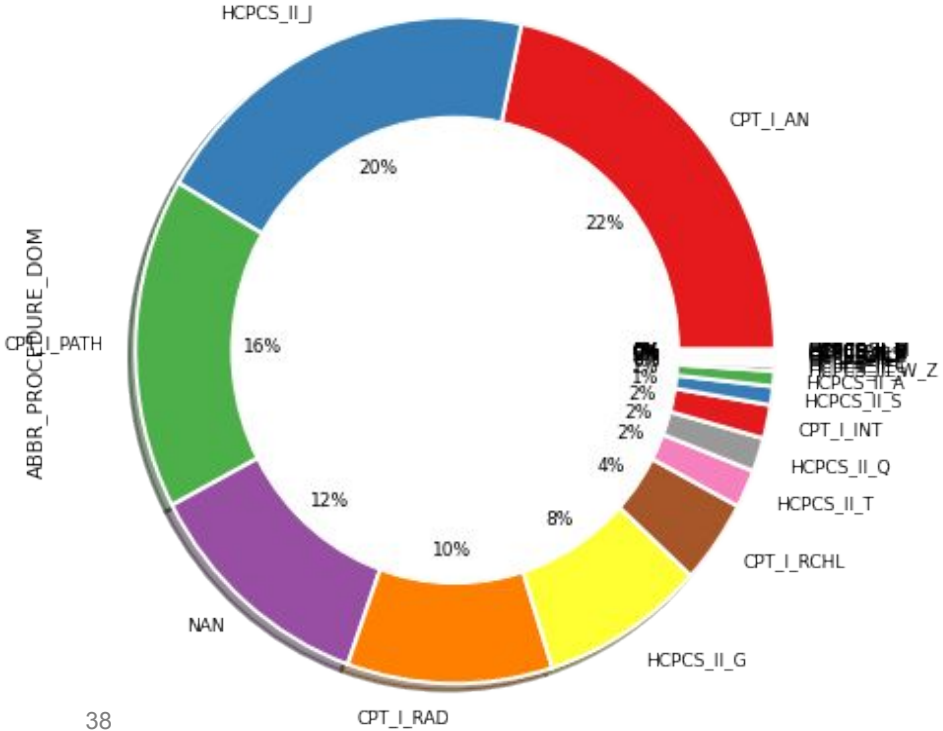
  - **BOOL_PROC_MOD**



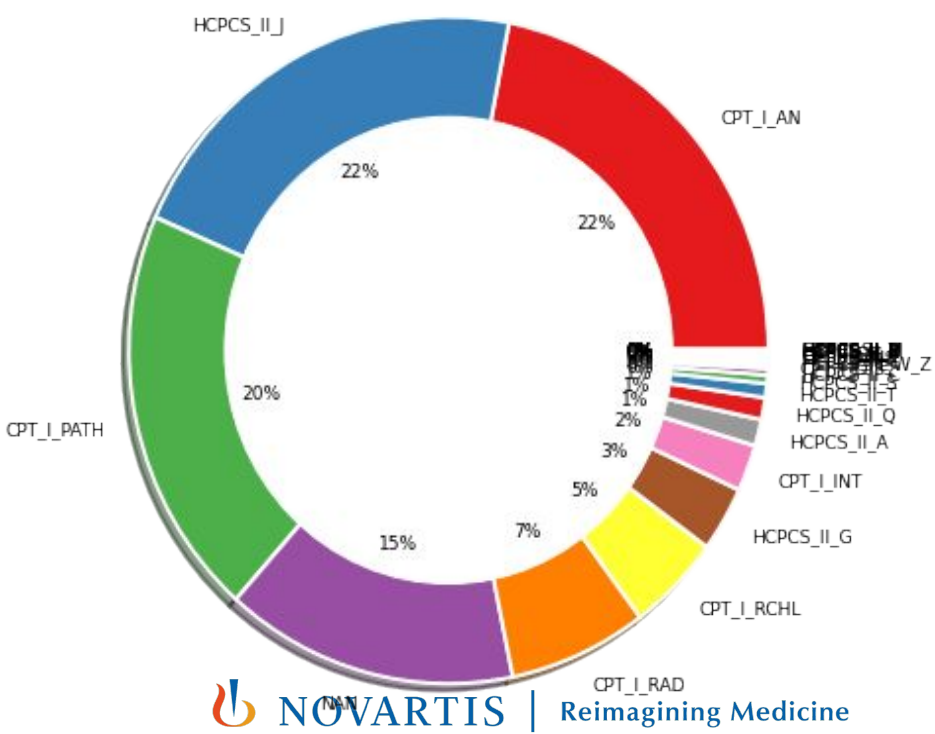
- Tracking each patient history till cancer diagnosis -
  - **WEEK_ID** -> ie the weekNumber_Year
  - Helps track the number of unique weeks the patient was under observation

EDA

Distribution of ABBR_PROCEDURE_DOM type for target==0



Distribution of ABBR_PROCEDURE_DOM type for target==1



Data Preparation

- Removed the factors which did not add any domain value to identify help in predicting the diagnosis
  - All CLAIM ID related columns
  - PATIENT_ID identifier

Data Splitting

- For each patient - post identification of the first change of treatment, all other data points were filtered out - avoid **information Leak.**
  - **only records with Service_date<=minimum 2nd line of treatment dates**
- Model training - split the data - Training : 75%, Testing : 25%
-

Patients w/o treatment change	7851
Patients w treatment change	7120
Total	**14971**

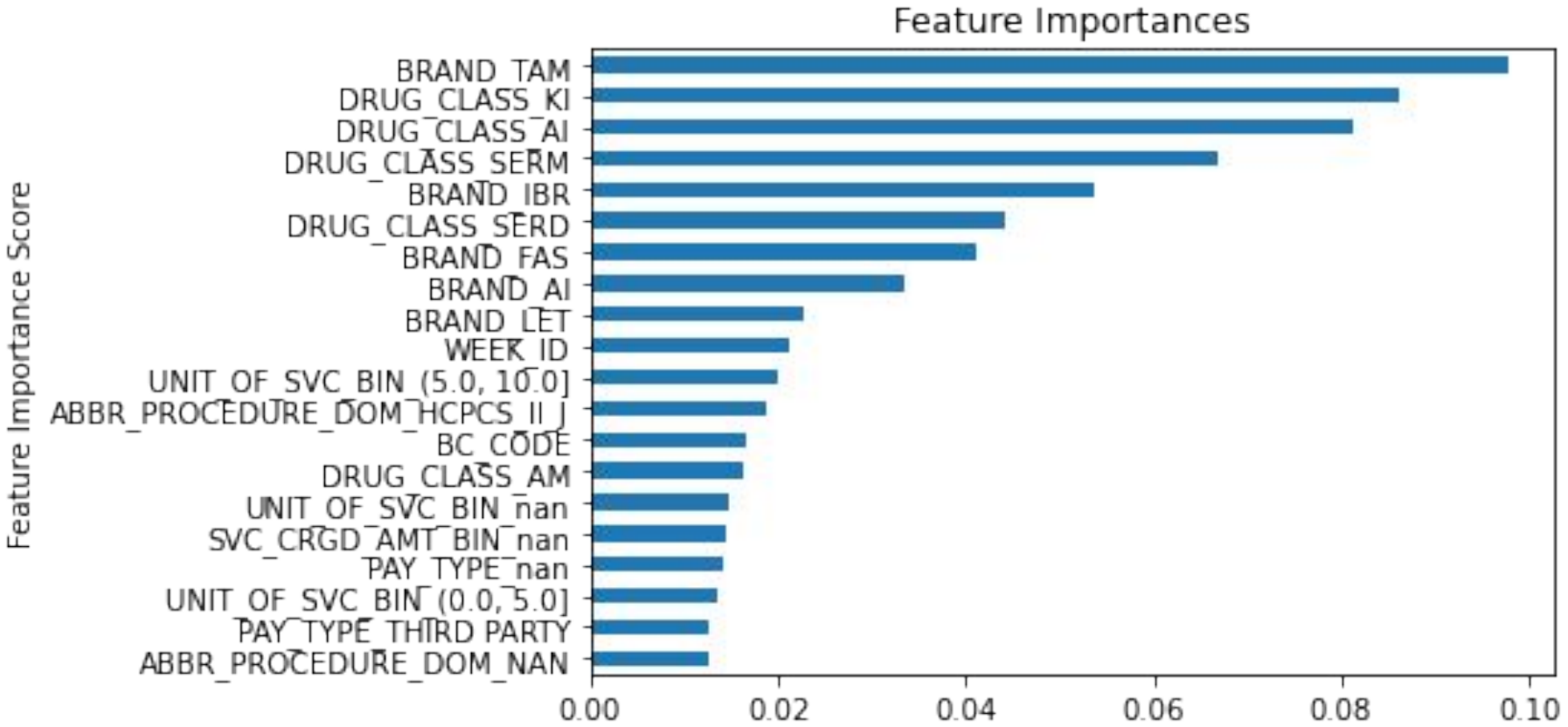
- Model training - Resampling data was not required since the 2 target categories have similar number of rows.

Model Training

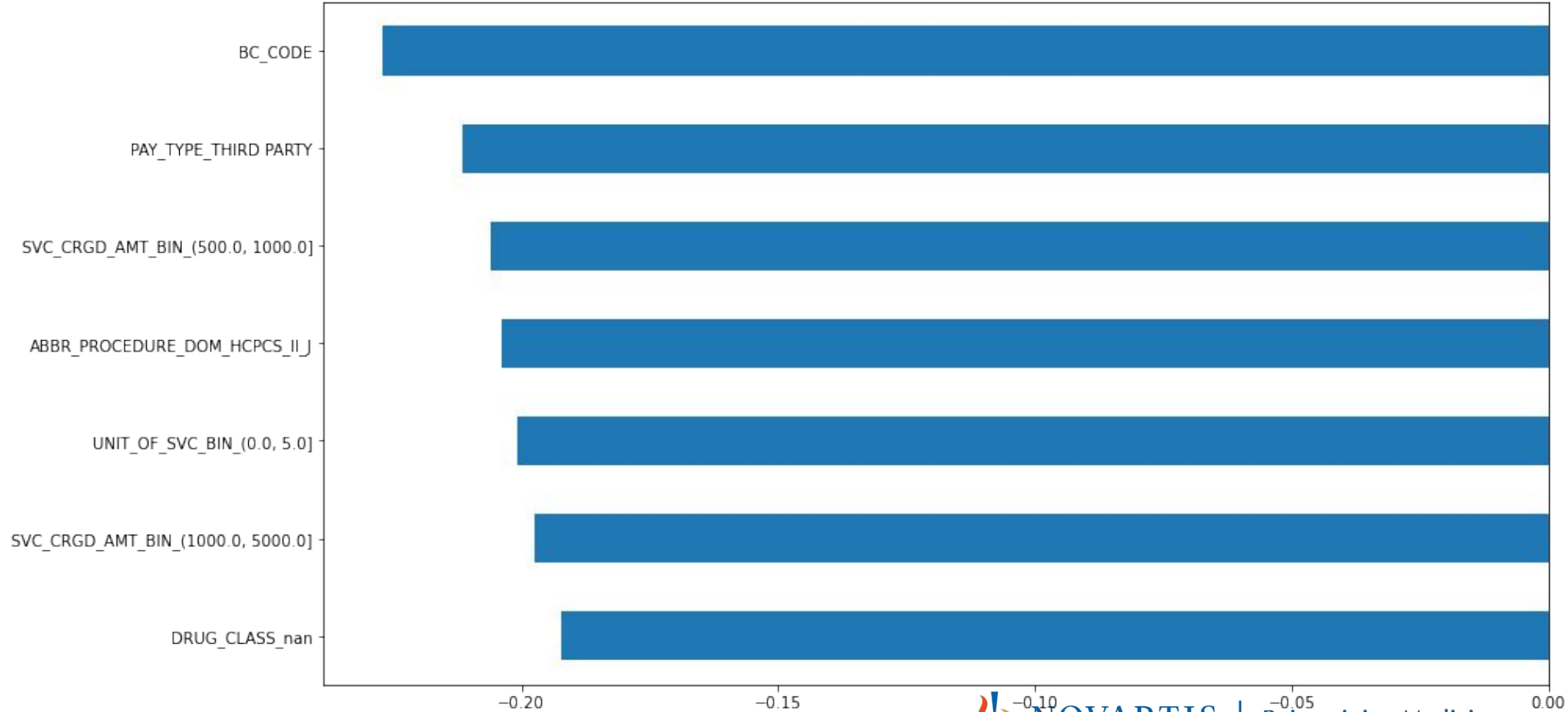
Model	Fitting time	Precision	Accuracy	Recall/Sensitivity	Specificity	F1_score	AUC_ROC
Decision Tree	0.422811	0.962499	0.962527	0.962408	0.962002	0.962526	0.963110
CatBoost	3.0612s	0.983712	0.971749	0.957048	0.961249	0.970197	0.971196
Random Forest	3.67735	0.954924	0.955180	0.955847	0.939804	0.955203	0.989384
Linear Discriminant Analysis	0.397495	0.688904	0.688999	0.689242	0.680963	0.689138	0.736073
K-Nearest Neighbors	0.352765	0.678210	0.678512	0.675929	0.737773	0.677464	0.720021
Logistic Regression	0.462069	0.669043	0.669026	0.669252	0.647479	0.669127	0.728146
Bayes	0.196315	0.613367	0.582593	0.593087	0.398044	0.562974	0.655229
Quadratic Discriminant Analysis	0.251577	0.582248	0.540978	0.541007	0.895786	0.457329	0.642296

- 8 different models
- 25% test data

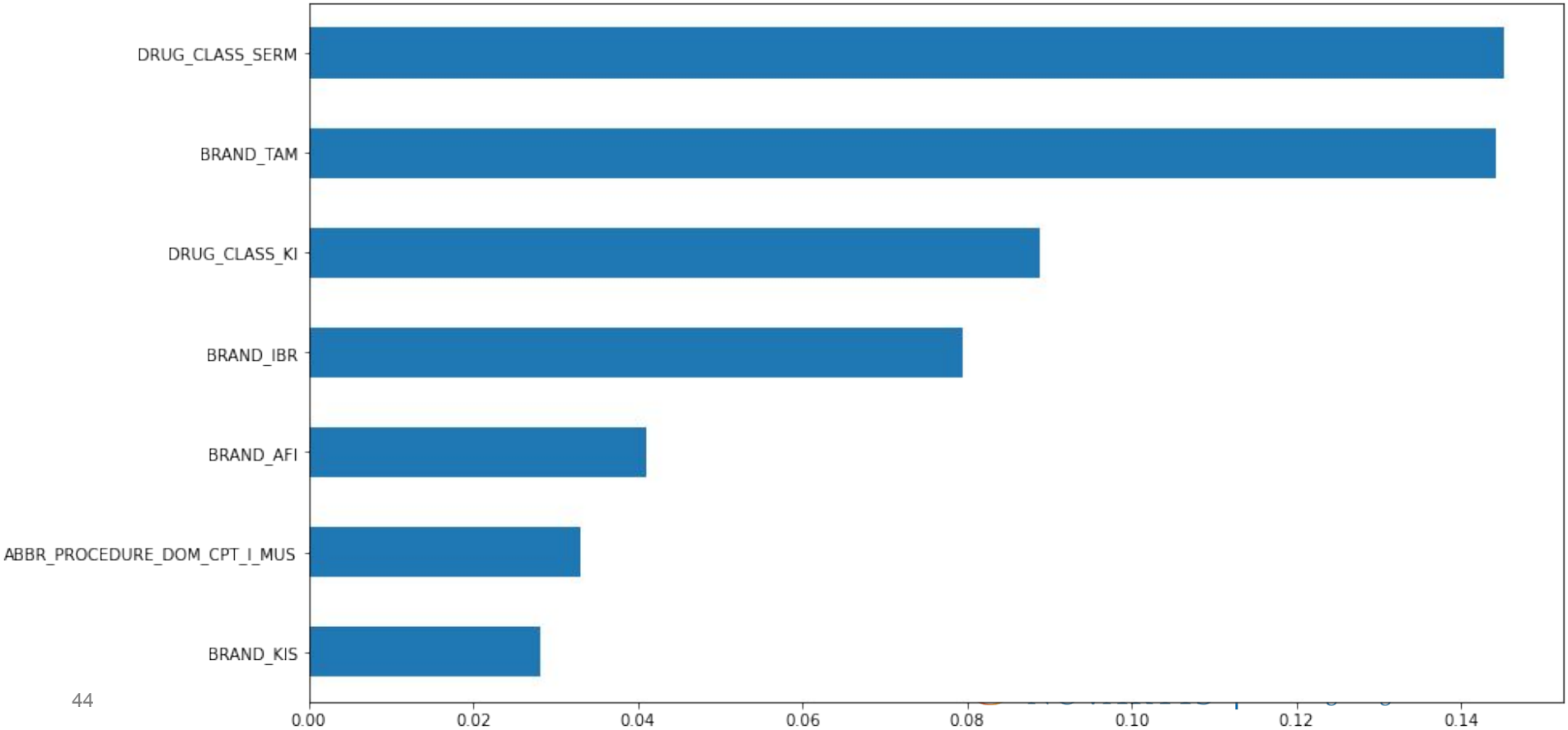
Top important features from the model



Top negatively correlated features



Top positively correlated features

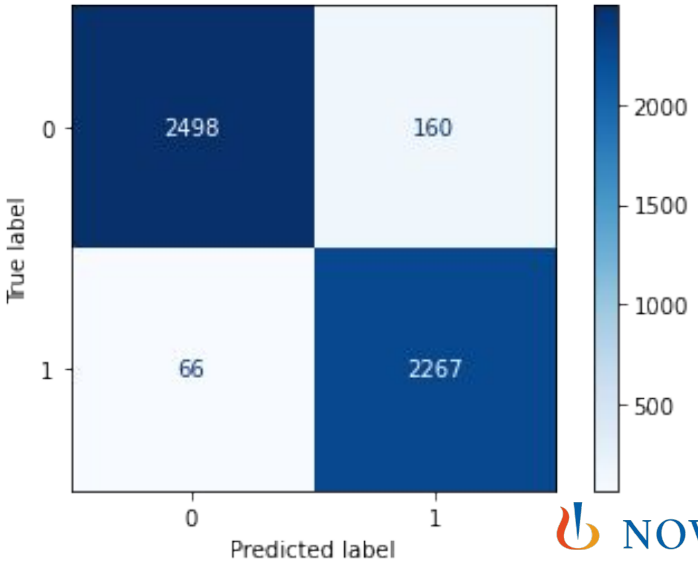


Results & Findings for Problem 2

- Decision Tree -> most consistent results.
- Training - 5 fold cross validation
- Top feature importances used by the model ->top 10 correlated columns.

Classification report.

- Accuracy: **0.9652**
- Sensitivity: **0.9624**
- Specificity: **0.9620**



Tools used

1. **Google cloud notebooks** to create a high capacity Jupyter notebook (needed more ram space, huge dataset, faster computing power)

2. **Kaggle kernel** for model training(cause of version maintainability, unmonitored running of code)

3. **Regex** for matching different drug codes and procedure codes.

4. **Python** modules like matplotlib and seaborn for visualization

Future works and applications

- Beyond this competition there is huge scope to increase the model's performance using hyper parameter tuning and/or stacking various models.
- Another approach we considered involved training a continual learning model which treats each patient separately and clubbed all the patients in training as one epoch.
  - This can be executed using a NN (maybe LSTM) or AutoML (requires research)
- Both the models currently developed take into account the domain knowledge and do not use any future information for prediction and do not require any more than 3 secs of training time on the free kaggle servers, reusing them or deploying them

Bibliography

https://union.ces.ncsu.edu/2019/10/male-breast-cancer-awareness/

https://www.nationalbreastcancer.org/breast-cancer-facts

https://blog.bonfire.com/cancer-ribbons/

https://allzonems.com/cpt-codes-the-three-categories-of-cpt-codes/

https://www.imaginis.com/breast-cancer-treatment/classes-of-breast-cancer-drugs-2

Q & A

NOVARTIS | Reimagining Medicine

THANK YOU