DSCI 552 Project

Presented by Group 13

Agenda

1. Introduction
2. Approach Discussion
3. Implementation Details
4. Summary and Q&A

Meet Our Group

Akhilaa

Sahithi

Mona

Ganesh

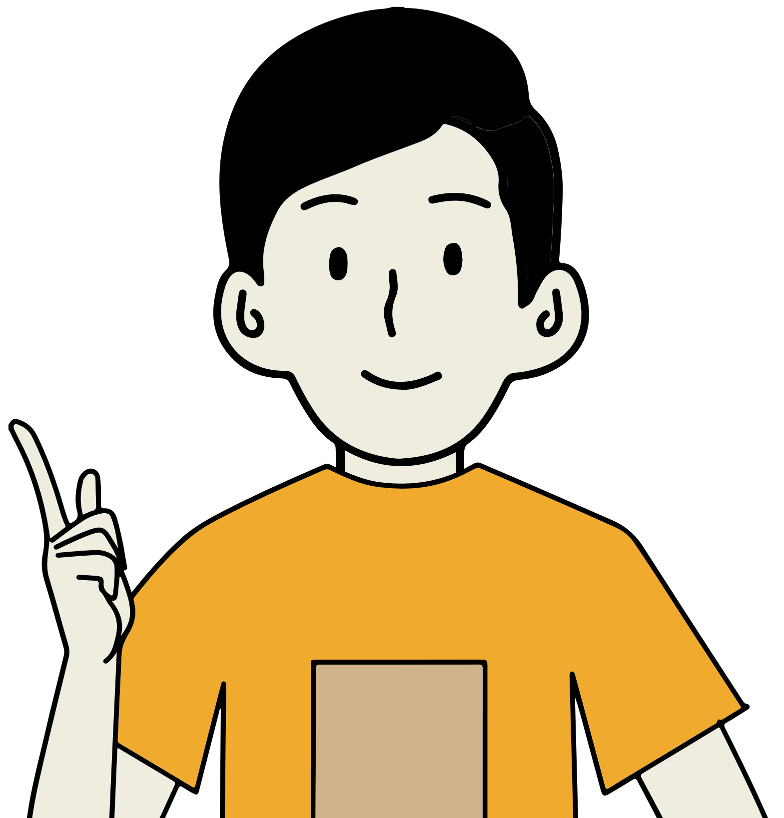
Introduction

Home Credit, established in 1997, aims to broaden financial inclusion by responsibly lending to individuals with limited credit history. By improving risk assessment, they seek to accept more loan applications and enhance the financial well-being of historically underserved populations.

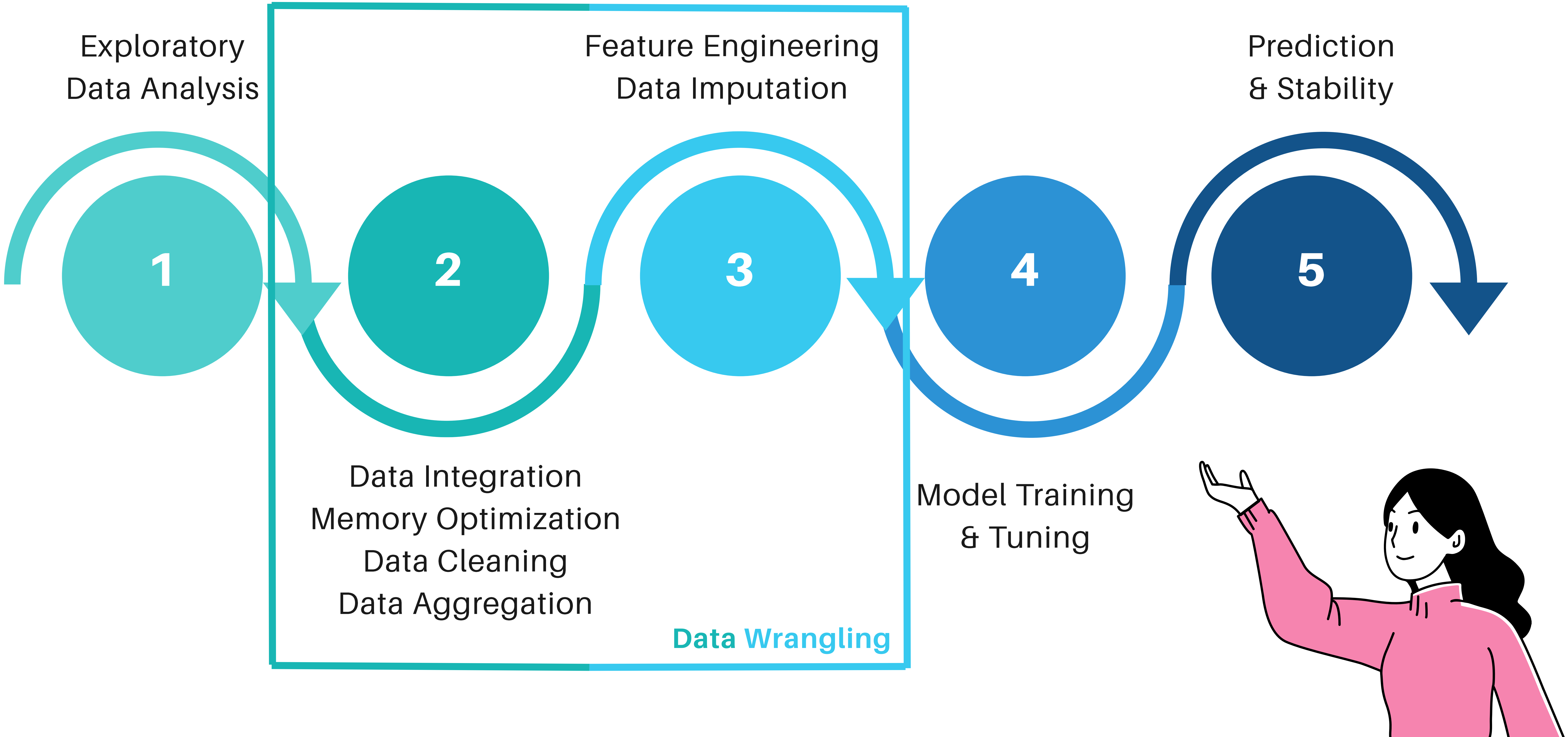
01 Credit History Challenges and Data Science Solutions

02 Understanding Score Calculation and Preserving Stability

03 Building a ML model that predicts the credit worthiness



PROCESS OVERVIEW



File 1

File 2

File 3

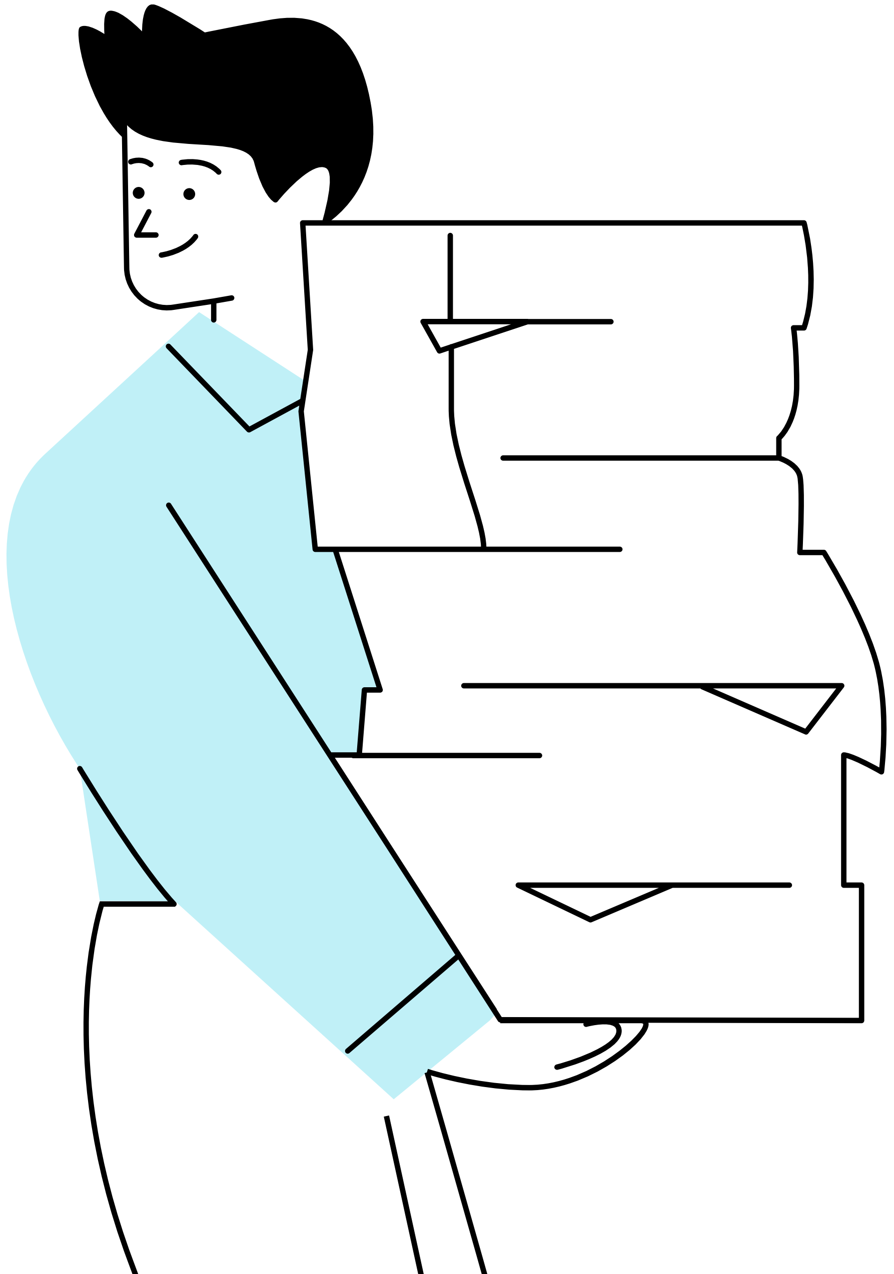
Given
Train Data
(Parquet format - 1.3 GB)

File 4

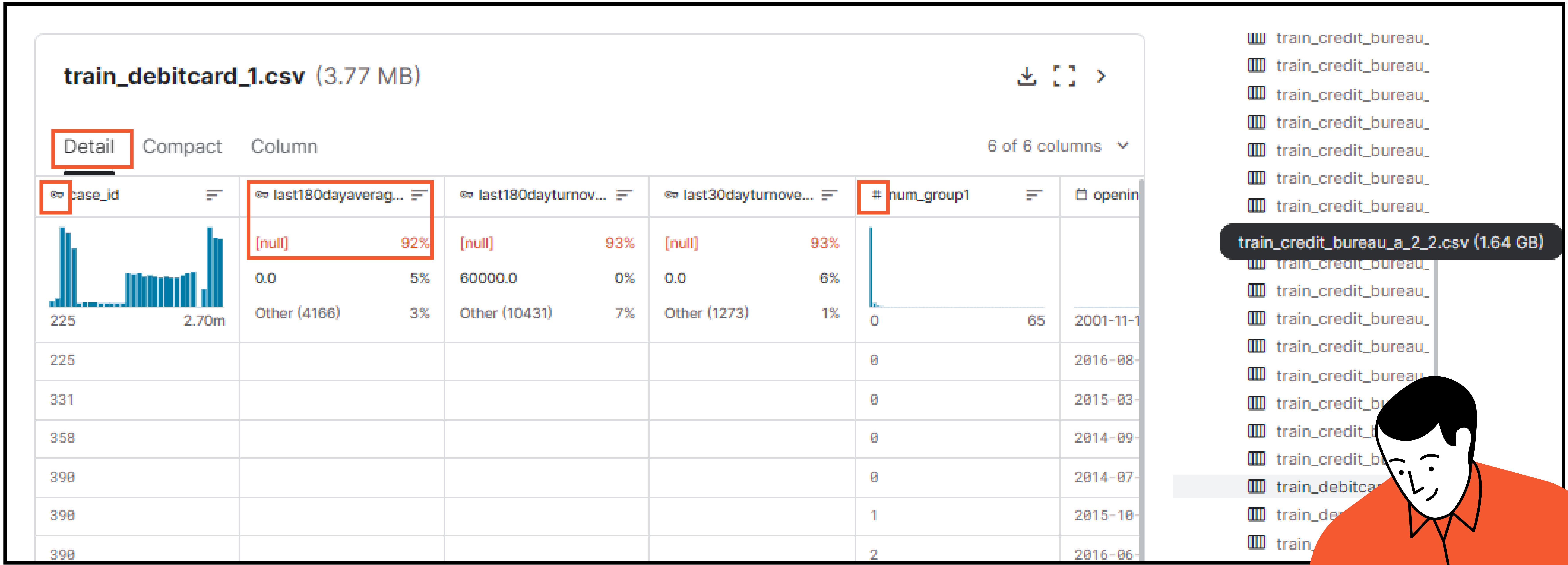
File 5

File 32

And one more thing... It takes many pandas to defeat one polar bear...

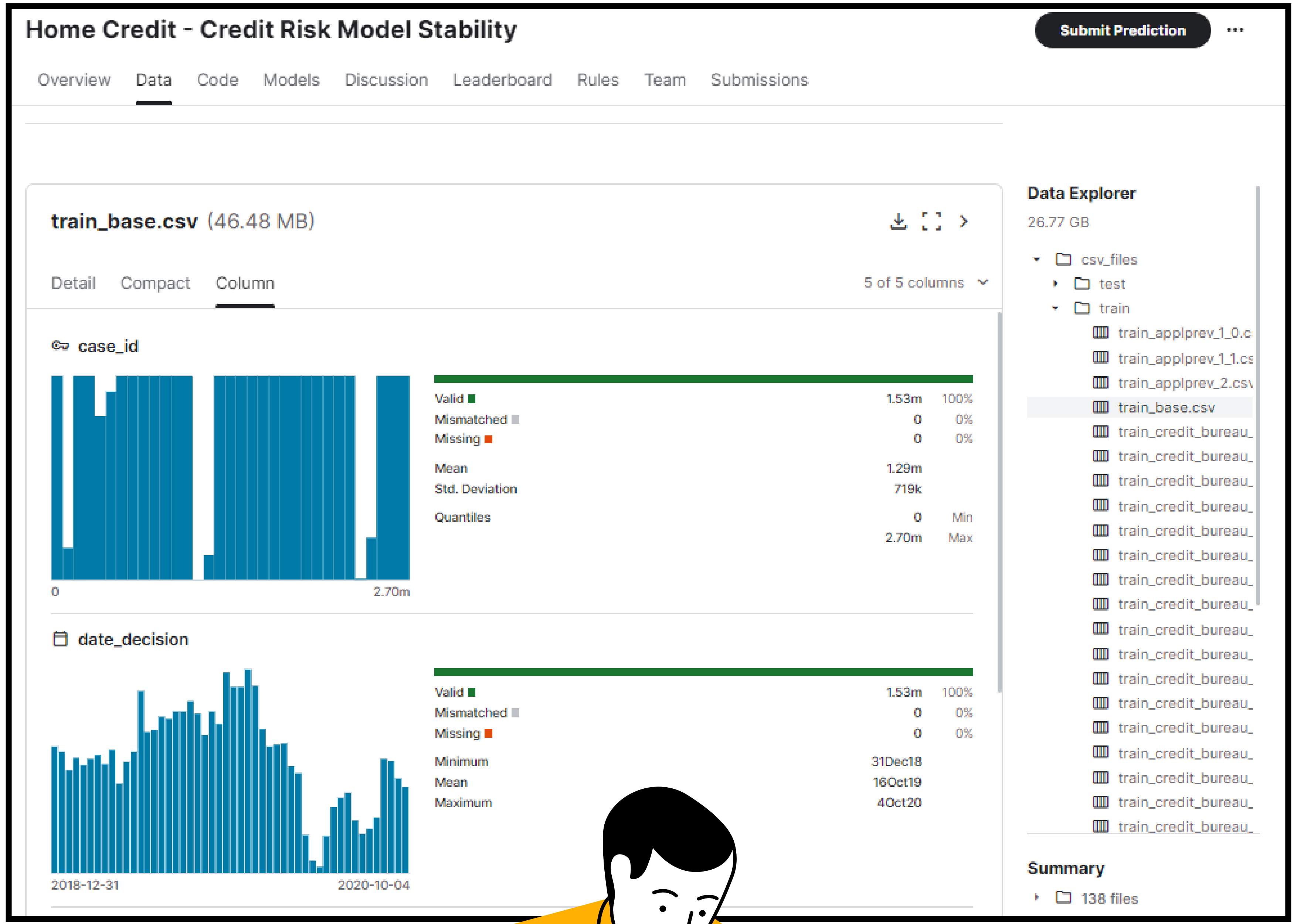


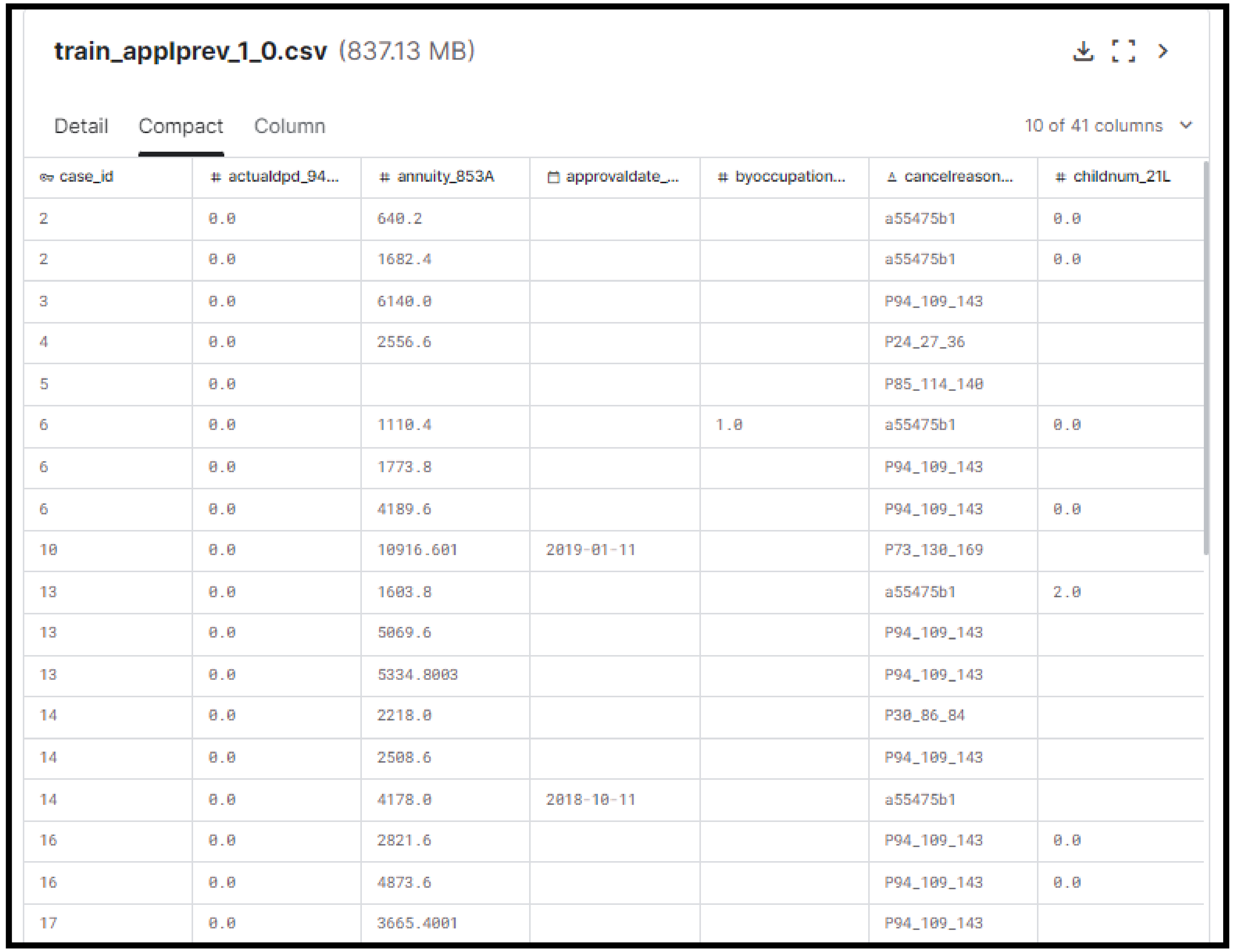
DATASET DESCRIPTION



Kaggle has a Data Explorer in the Data section which serves as a great starter guide to understand the data structure and distribution

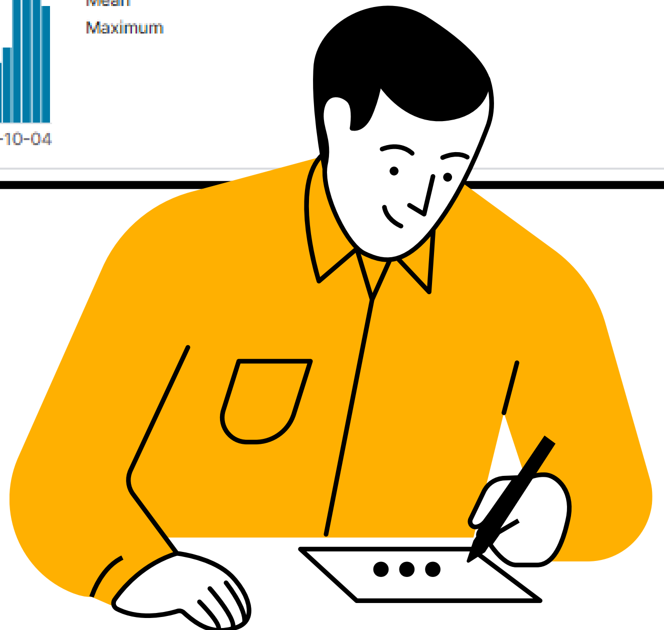
DATASET DESCRIPTION CONTINUED…



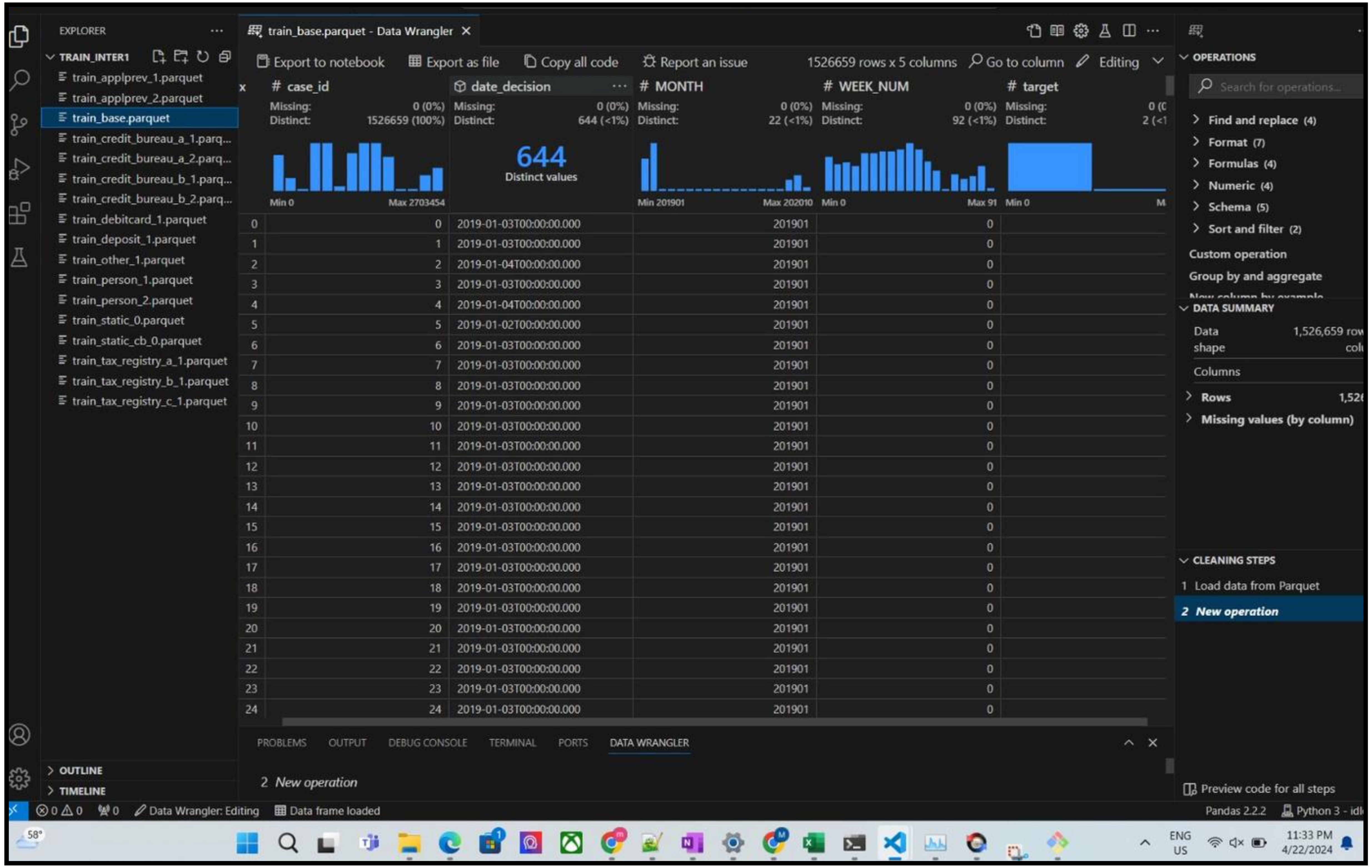

train_applprev_1_0.csv (837.13 MB)

Detail Compact Column 10 of 41 columns

case_id	actualdpd_94...	annuity_853A	approvaldate_...	byoccupation...	cancelreason...	childnum_21L
2	0.0	640.2			a55475b1	0.0
2	0.0	1682.4			a55475b1	0.0
3	0.0	6140.0			P94_109_143	
4	0.0	2556.6			P24_27_36	
5	0.0				P85_114_140	
6	0.0	1110.4		1.0	a55475b1	0.0
6	0.0	1773.8			P94_109_143	
6	0.0	4189.6			P94_109_143	0.0
10	0.0	10916.601	2019-01-11		P73_130_169	
13	0.0	1603.8			a55475b1	2.0
13	0.0	5069.6			P94_109_143	
13	0.0	5334.8003			P94_109_143	
14	0.0	2218.0			P30_86_84	
14	0.0	2508.6			P94_109_143	
14	0.0	4178.0	2018-10-11		a55475b1	
16	0.0	2821.6			P94_109_143	0.0
16	0.0	4873.6			P94_109_143	0.0
17	0.0	3665.4001			P94_109_143	



We can explore the detailed insights using Column and Compact tabs…

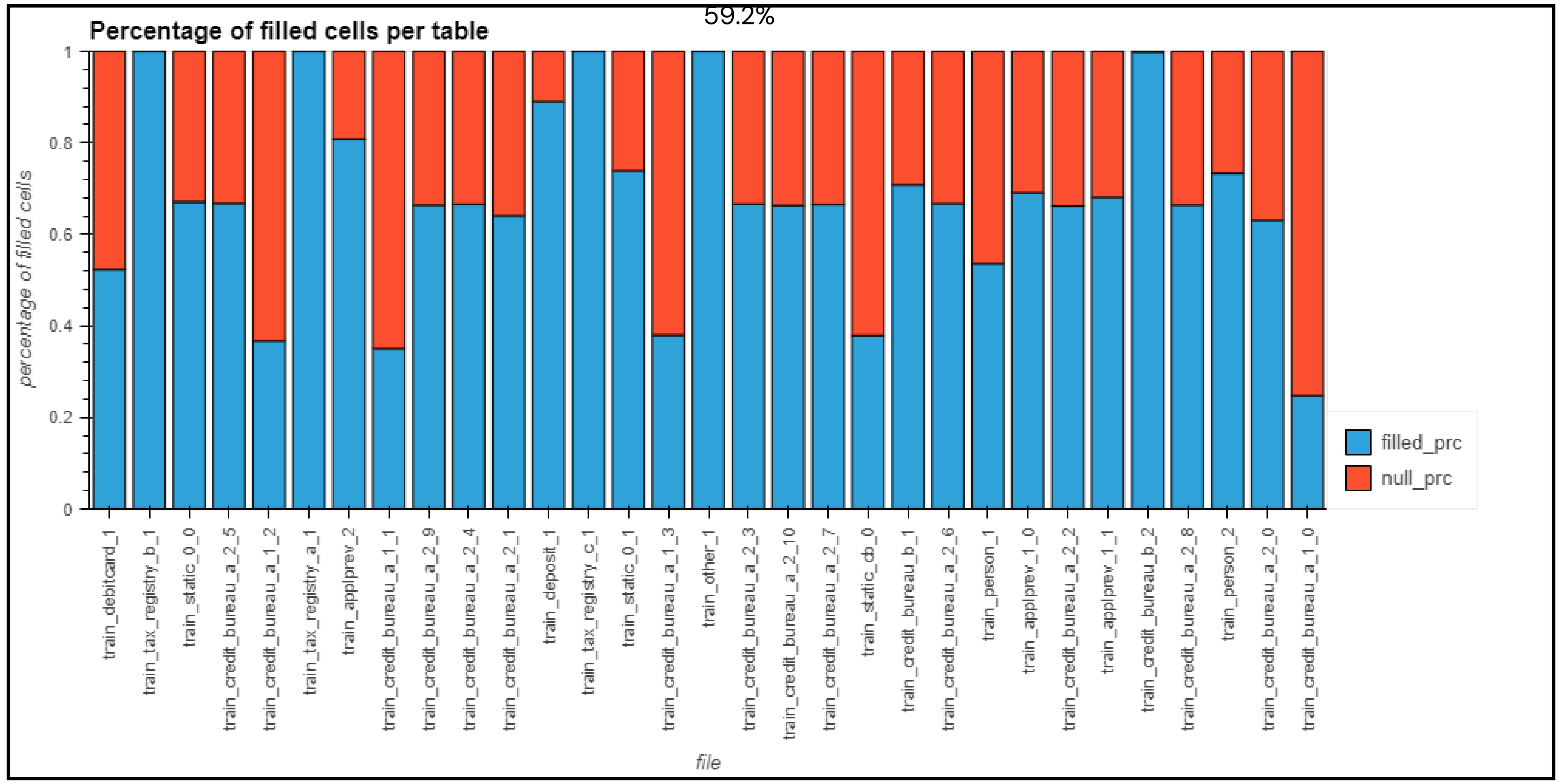


Exploratory Data Analysis

We used the Data Wrangler Extension in MS VS Code to drill down and gain deeper insights of the data

Exploratory Data Analysis Continued...

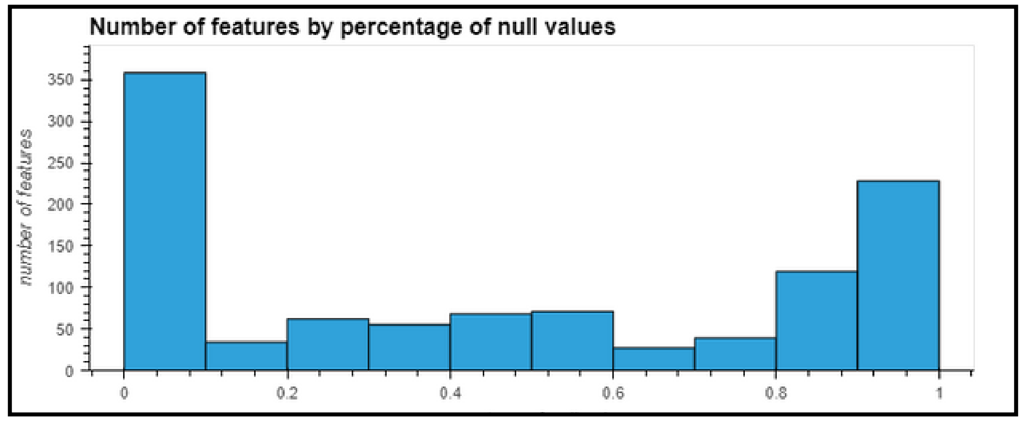
Null cells - 40.8%, Filled cells - 59.2%

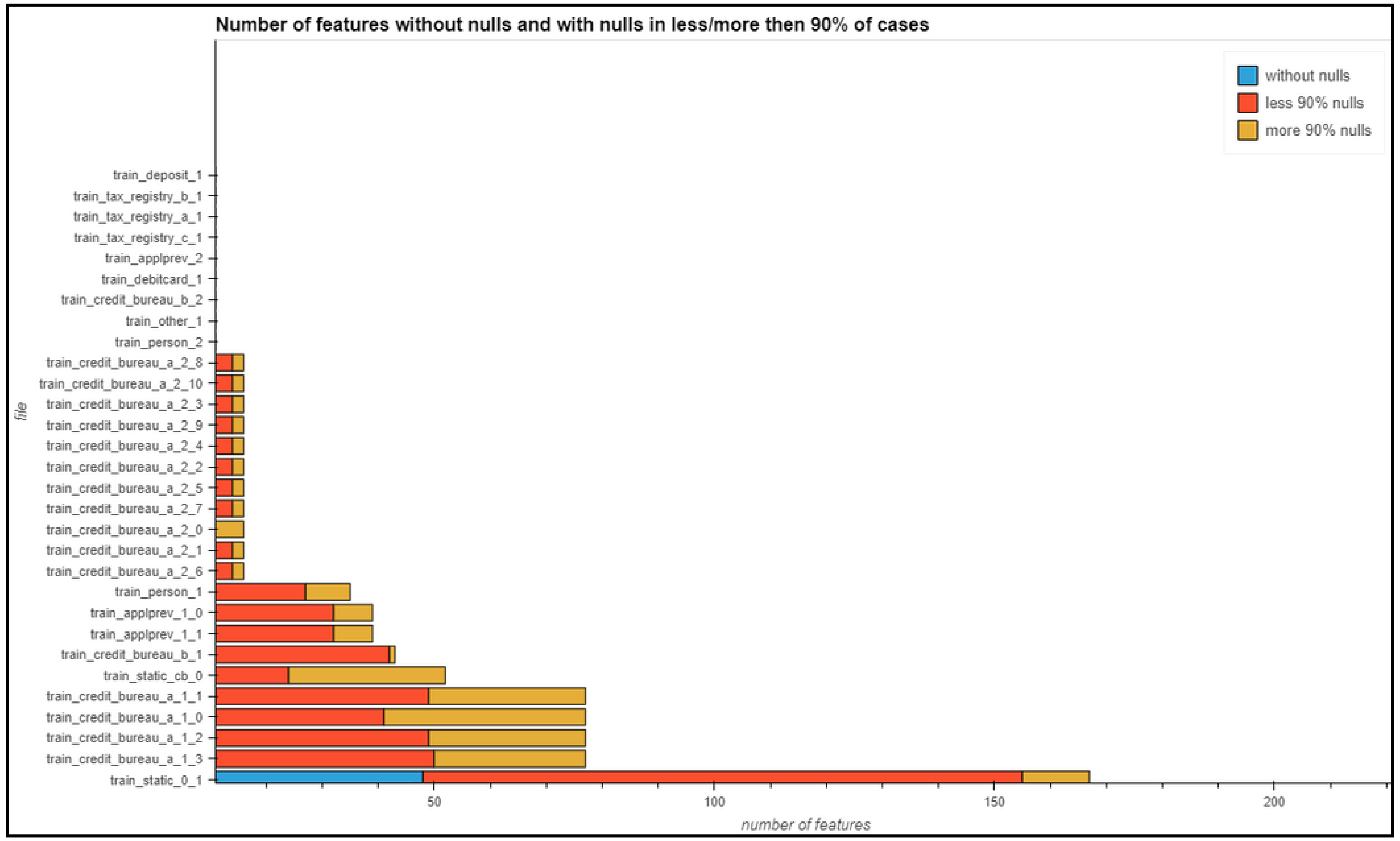


Variable	Description
465 unique values	436 unique values

The competition sponsor divided the features into custom types:

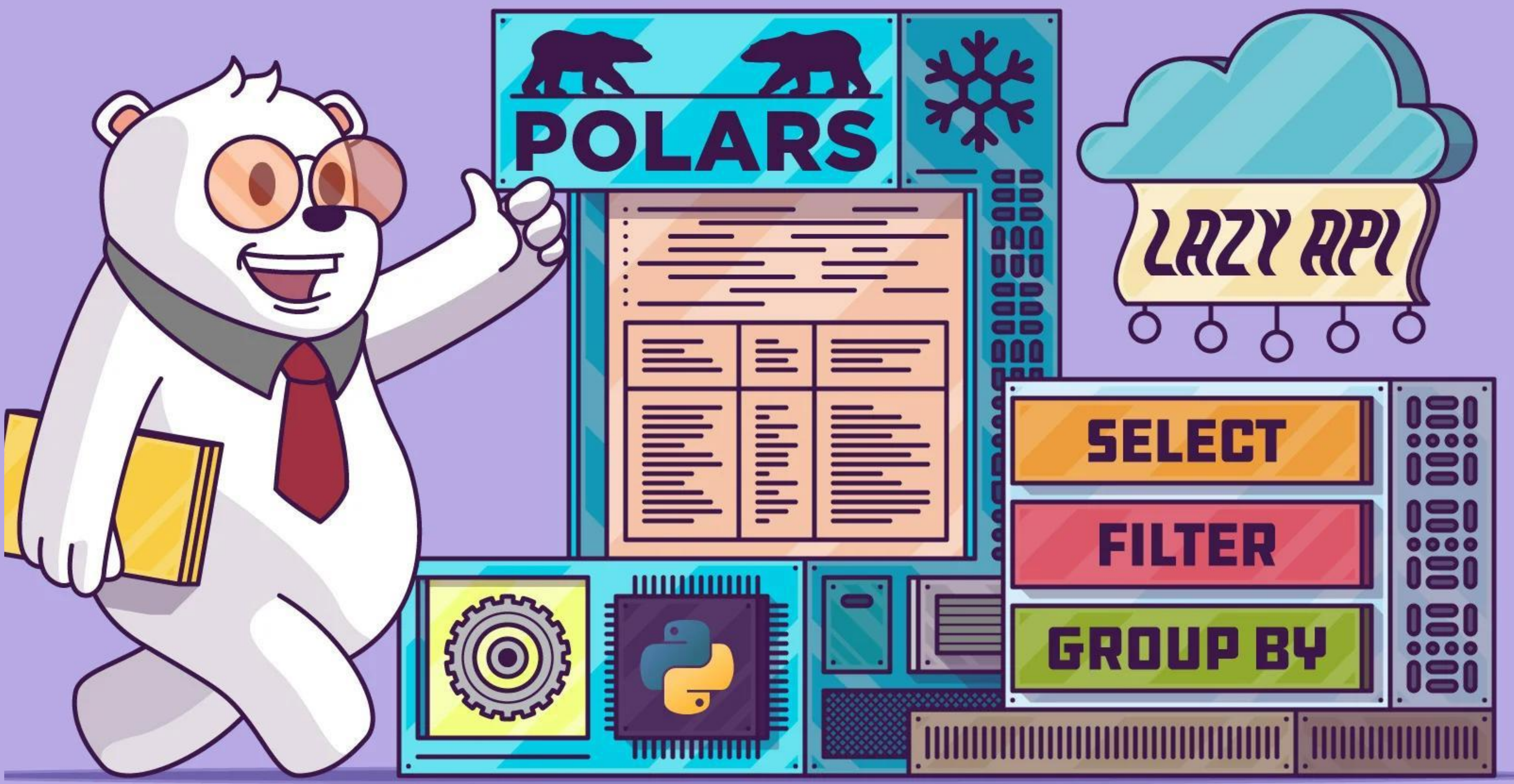
- **P** - Transform DPD (Days past due)
- **M** - Masking categories
- **A** - Transform amount
- **D** - Transform date
- **T** - Unspecified Transform
- **L** - Unspecified Transform





- We have used .describe() to explore the distribution of all the features.
- By analyzing the min, max, avg and inter quartile regions for the numeric features we gained a lot of insights into the data.
- We have used the line plots for numeric data and bar plots for categorical data (stacked bar charts) to visualize the variation,
- After the detailed inspection and examination of data we move on to the Data Integration and Memory Optimization part.

Data Integration (DI) & Memory Optimization (MO)



Read the Data using Polars

Optimize memory by strategic data transformation (Ex: Int64 to Int8)

Combine the multiple instances of same file types and use GroupBy

DI & MO

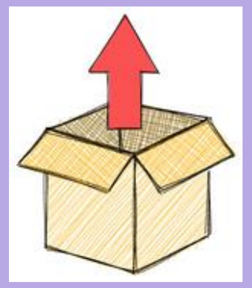
Initial Train Data

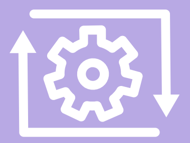
File 1

File 2

...

File 32





Transformed Train Data

File 1

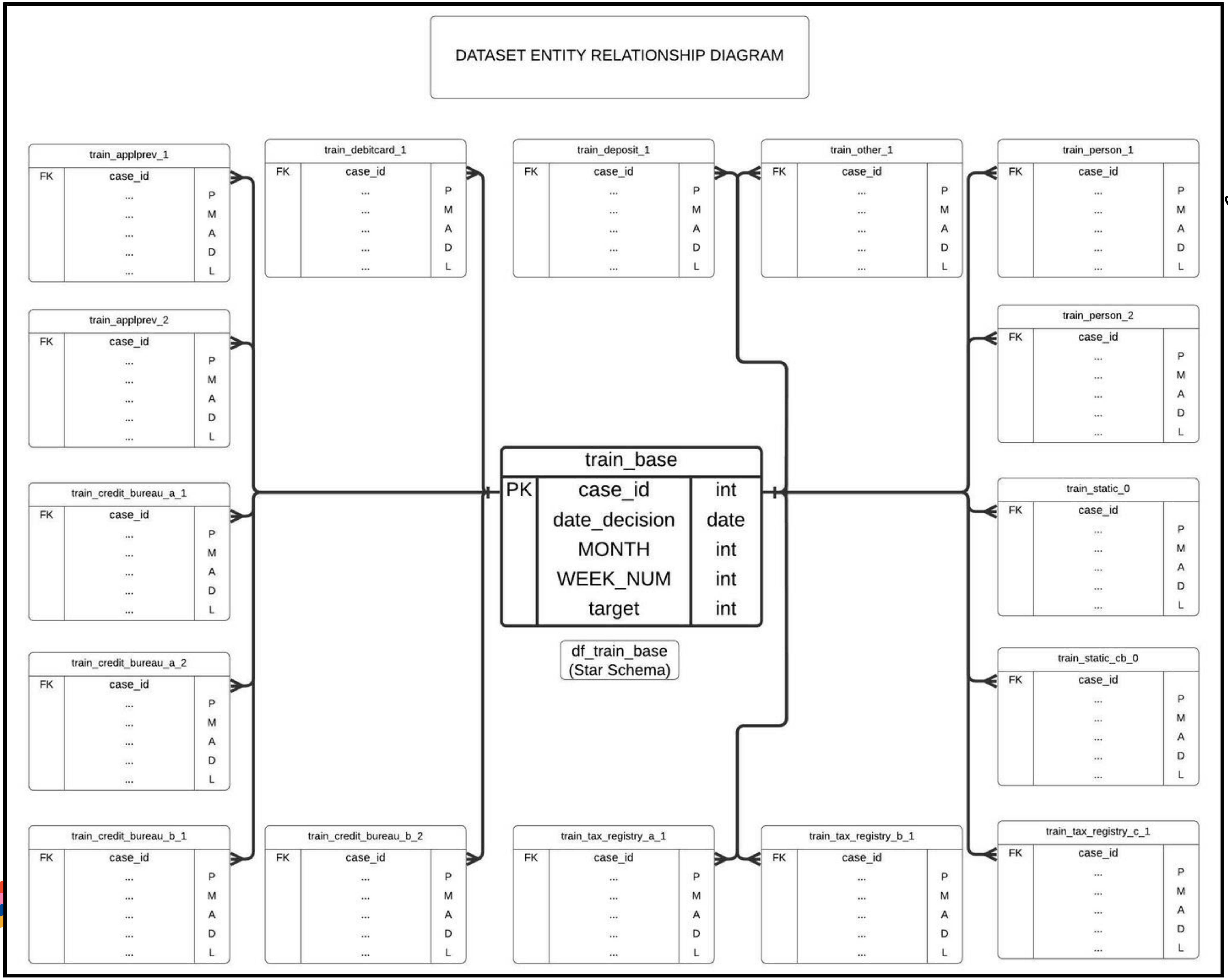
File 2

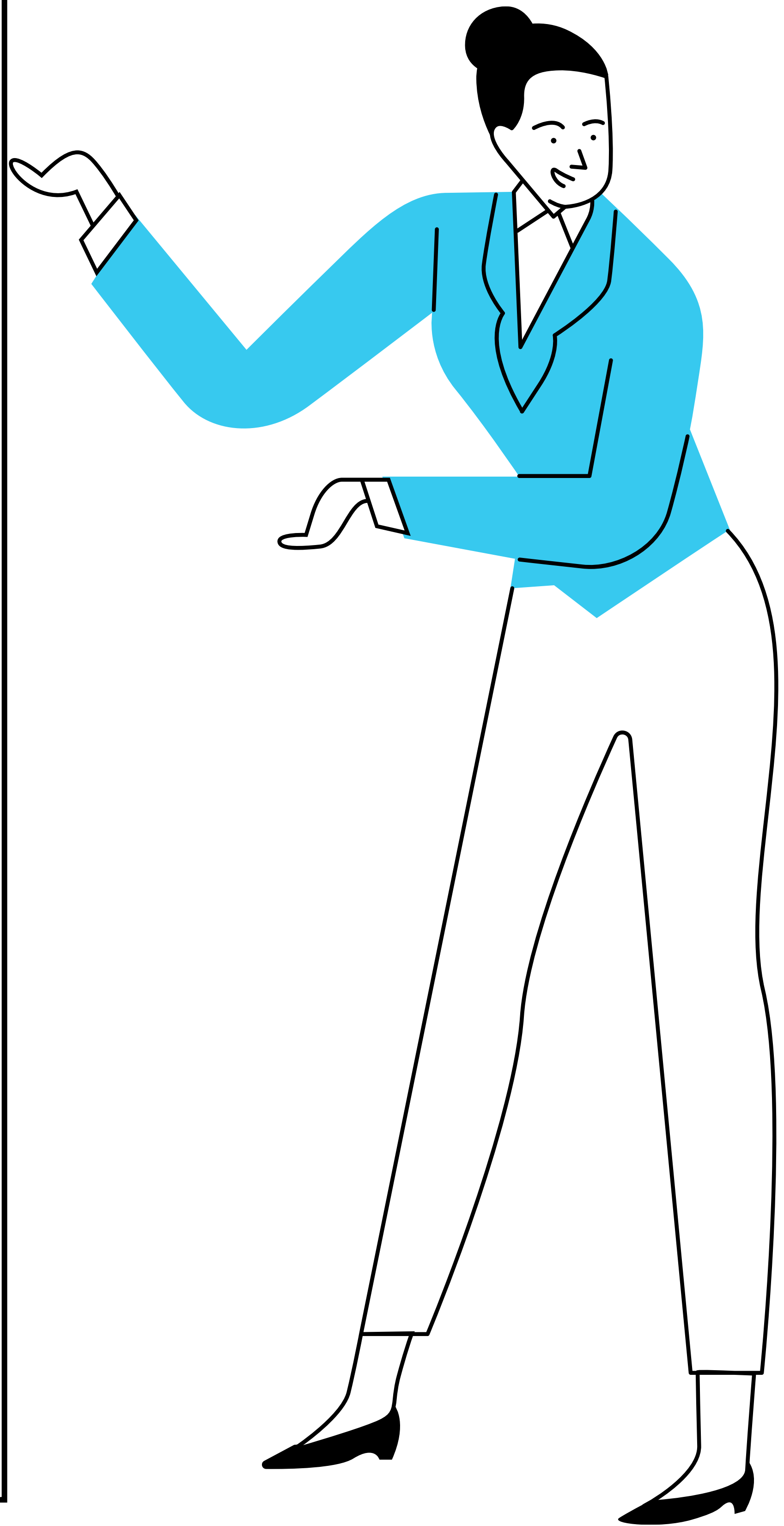
...

File 17

DATASET ENTITY RELATIONSHIP DIAGRAM

Our given train data has no duplicate column names in the total dataset





Data Cleaning & Data Aggregation

Drop the Columns or features that have null values > threshold

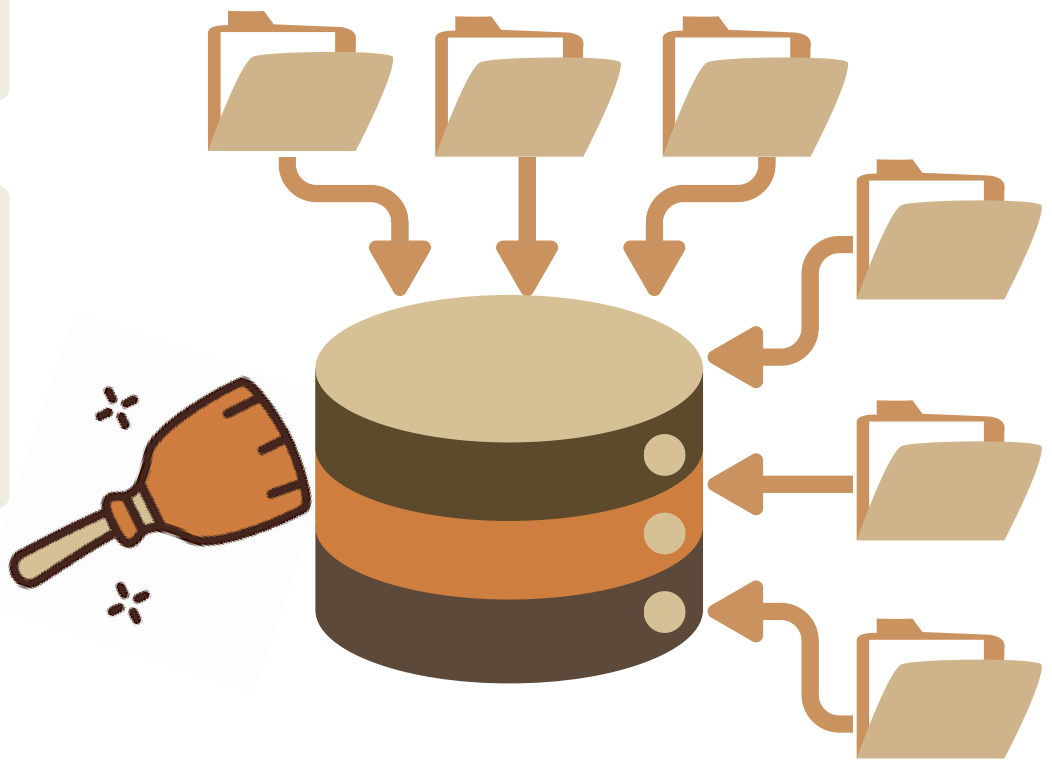
For Categorical Columns drop columns where value frequency is equal to 1 or > 180

Select the Train Base (df_train_base) as our main file to add other features

Now we will aggregate the data from the other 16 files on to Train Base using Case_ID as our key

After Data Aggregation we end up with a single file (df_base_train) :

- No duplicate Case_IDs
- No duplicate Columns (features)
- Still has missing values



Feature Engineering & Data Imputation

Add Decision_Month and Decision_Week columns to our df_train_base which later helps us to convert out date type attributes to days

Group the columns by correlation so that we end with groups of columns having similar data relationships

Use reduce_groups function to select most representative column within the group (Dimensionality Reduction)

Data Imputation – Fill the null values of numeric data with the mean and categorical data with mode (if not "null")

Converting the Processed Data to Pandas

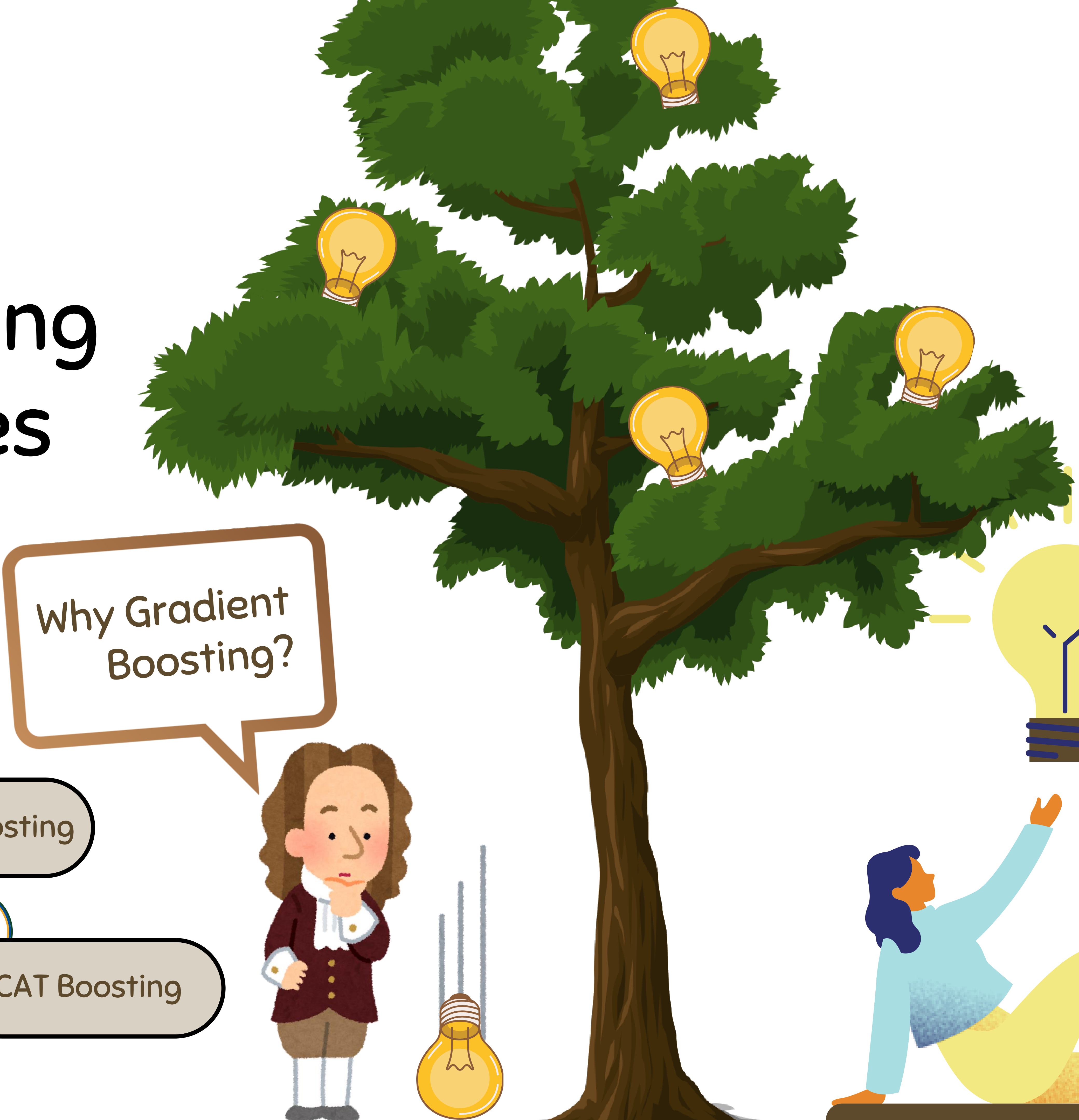
Convert the final df_train_base (filled) that we have after Data Imputation to Pandas as it has a good ecosystem support

This Dataframe df_train_base in pandas format will be now fed into model training as input

This marks the end to the Data Wrangling Phase...

Pandas has better integration support and with Scikit Learn, TensorFlow, XGBoost, Seaborn and other python libraries and frameworks. Hence, it offers a wide range of choices to the user to build a robust model utilizing best libraries

Modeling Choices



Random Forest

Gradient Boosting

Light GBM

CAT Boosting

LightGBM CatBoost Random Forest

Modeling Results

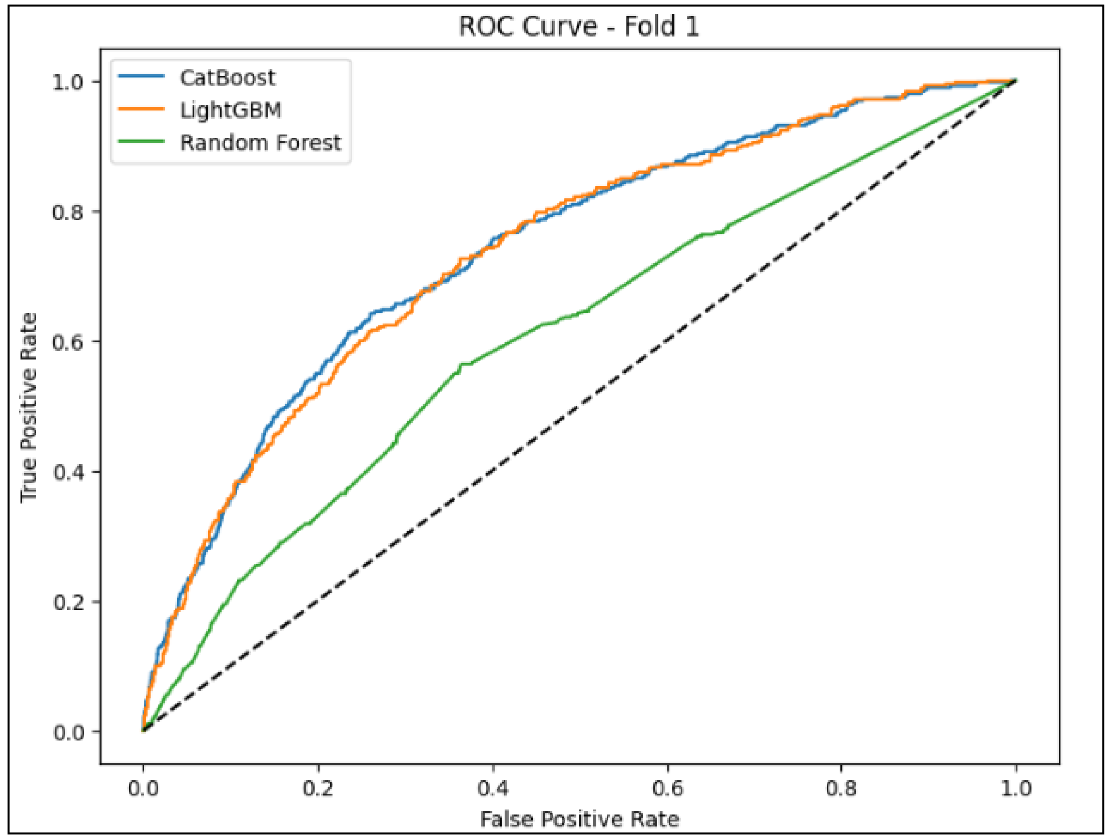
We tried adjusting different parameters for these models and the best results for each model is shown in the table

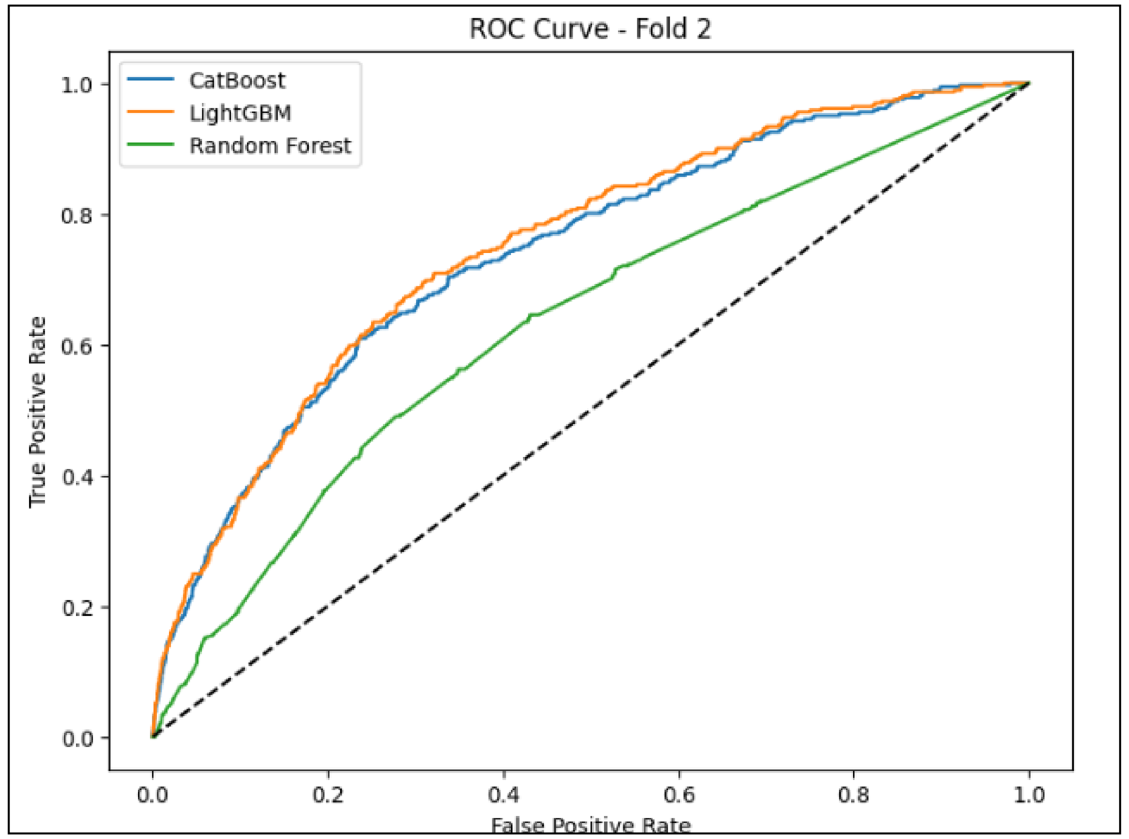
Light GBM (Best)	CAT Boost (Best)	Random Forest (Best)
MAX AUC SCORE (0.7861)	MAX AUC SCORE: (0.7747)	MAX AUC SCORE (0.6673)

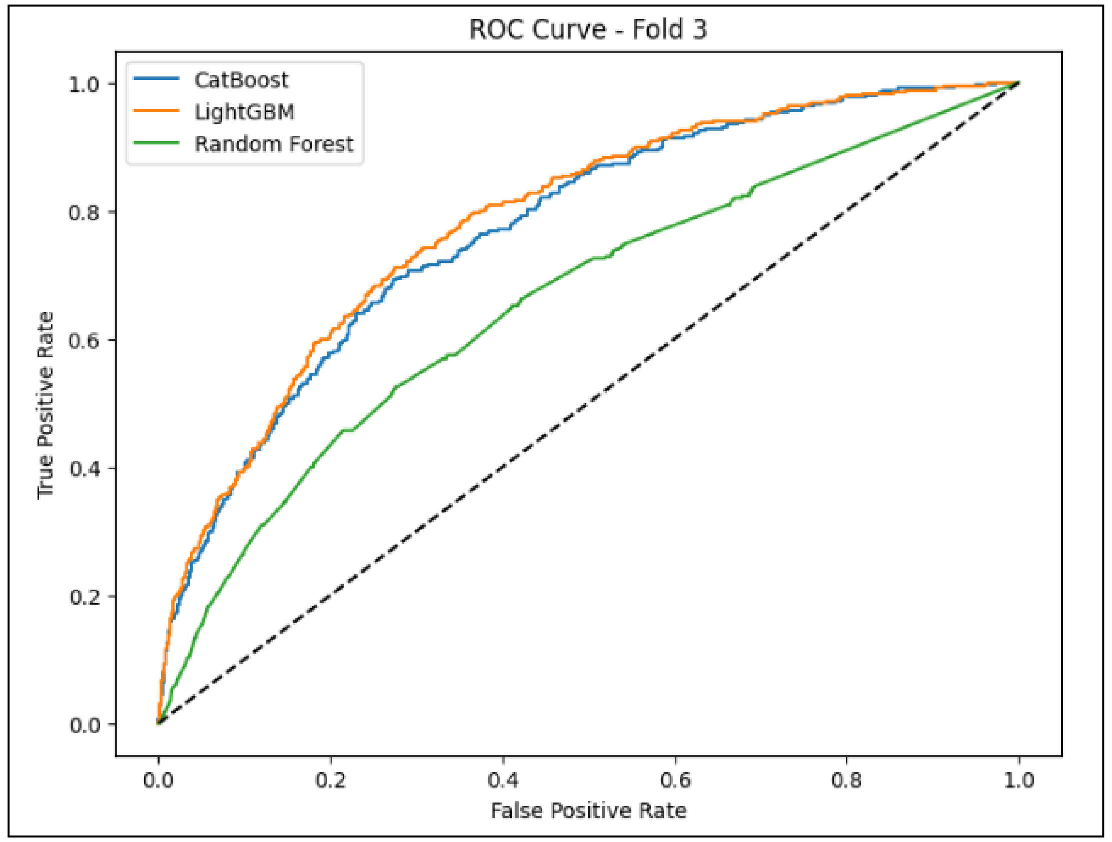
Kaggle Stability Criteria (Gini Stability)

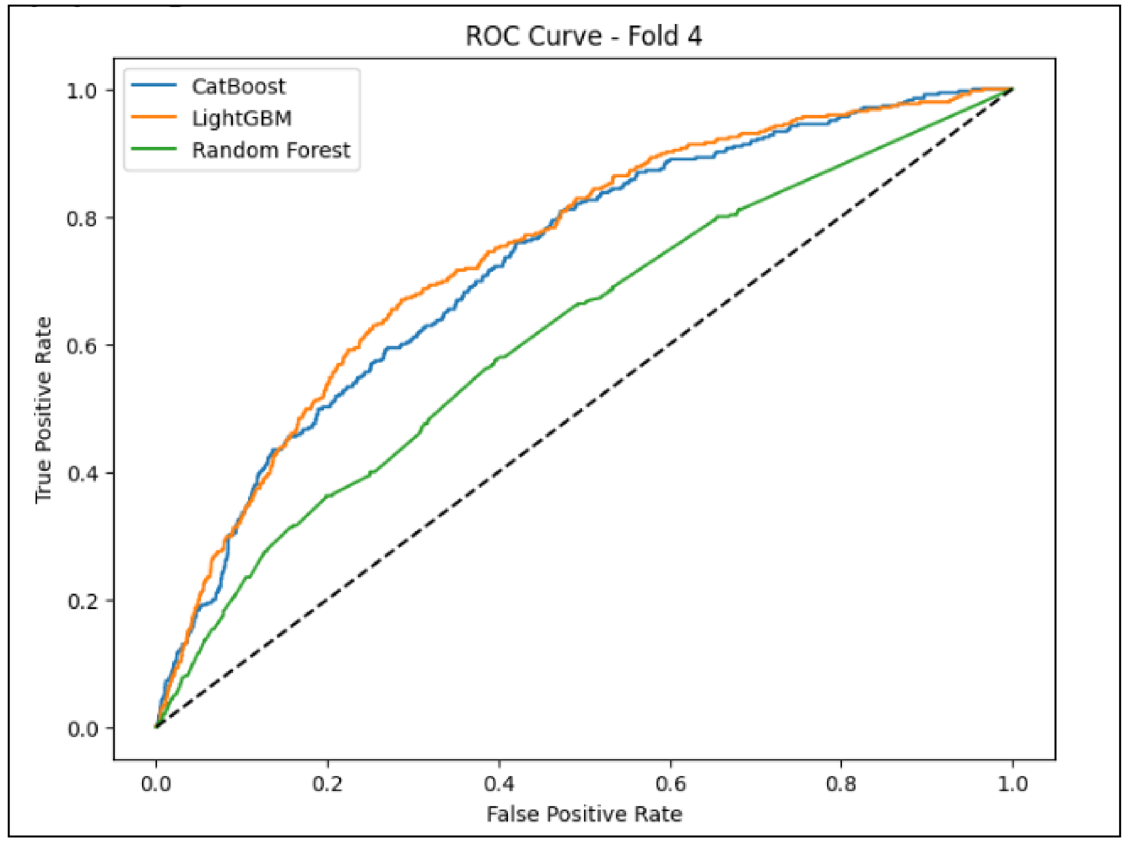
LightGBM CatBoost Random Forest

Modeling Results Continued...









Ablation Study

Type	Conditions	Light GBM	CAT Boost	Random Forest
DW	NULL FILTERING FREQUENCY THESHOLD: 0.50	MAX AUC SCORE:0.7701	MAX AUC SCORE: 0.7624	MAX AUC SCORE: 0.6102
DW	NULL FILTERING FREQUENCY THESHOLD: 0.70	MAX AUC SCORE: 0.7844	MAX AUC SCORE: 0.7716	MAX AUC SCORE: 0.6573
DW	NULL FILTERING FREQUENCY THESHOLD: 0.80	MAX AUC SCORE:0.7820	MAX AUC SCORE: 0.7733	MAX AUC SCORE: 0.6495

DW = Examining Data Wrangling Steps

Ablation Study Continued...

Type	Conditions	Light GBM	CAT Boost	Random Forest
DW	CATEGORICAL FREQUENCY THESHOLD: 160	MAX AUC SCORE: 0.7856	MAX AUC SCORE: 0.7736	MAX AUC SCORE: 0.6523
DW	CATEGORICAL FREQUENCY THESHOLD: 180 (HIGH AVG & STABILITY)	MAX AUC SCORE: 0.7844	MAX AUC SCORE: 0.7716	MAX AUC SCORE: 0.6573
DW	CATEGORICAL FREQUENCY THESHOLD: 200	MAX AUC SCORE: 0.7834	MAX AUC SCORE: 0.7747	MAX AUC SCORE: 0.6429

DW = Examining Data Wrangling Steps

Ablation Study Continued...

Type	Conditions	Light GBM	CAT Boost	Random Forest
ML	LGB ESTIMATORS: 1000 CAT ESTIMATORS: 4000 RF ESTIMATORS: 60	MAX AUC SCORE: 0.7841	MAX AUC SCORE: 0.7674	MAX AUC SCORE: 0.6557
ML	LGB ESTIMATORS: 1500 CAT ESTIMATORS: 5000 RF ESTIMATORS: 80	MAX AUC SCORE: 0.7844	MAX AUC SCORE: 0.7714	MAX AUC SCORE: 0.6605
ML	LGB ESTIMATORS: 2000 CAT ESTIMATORS: 6000 RF ESTIMATORS: 100	MAX AUC SCORE: 0.7861	MAX AUC SCORE: 0.7716	MAX AUC SCORE: 0.6673

ML = Examining Modeling Steps

Key Accomplishments

#	Team	Members	Score	Entries	Last	Join
419	ML_DS_GROUP_13		0.584	5	21h	

Future Scope & Conclusion

For Data Imputation we are planning to use K-Nearest Neighbors (KNN) Imputation to try improving our missing values filling strategy

Try to Optimize strategies for handling categorical data

Try to use more Ensemble Methods

XG Boosting?

Q & A

Really appreciate any suggestions for future implementation...

Thank You!