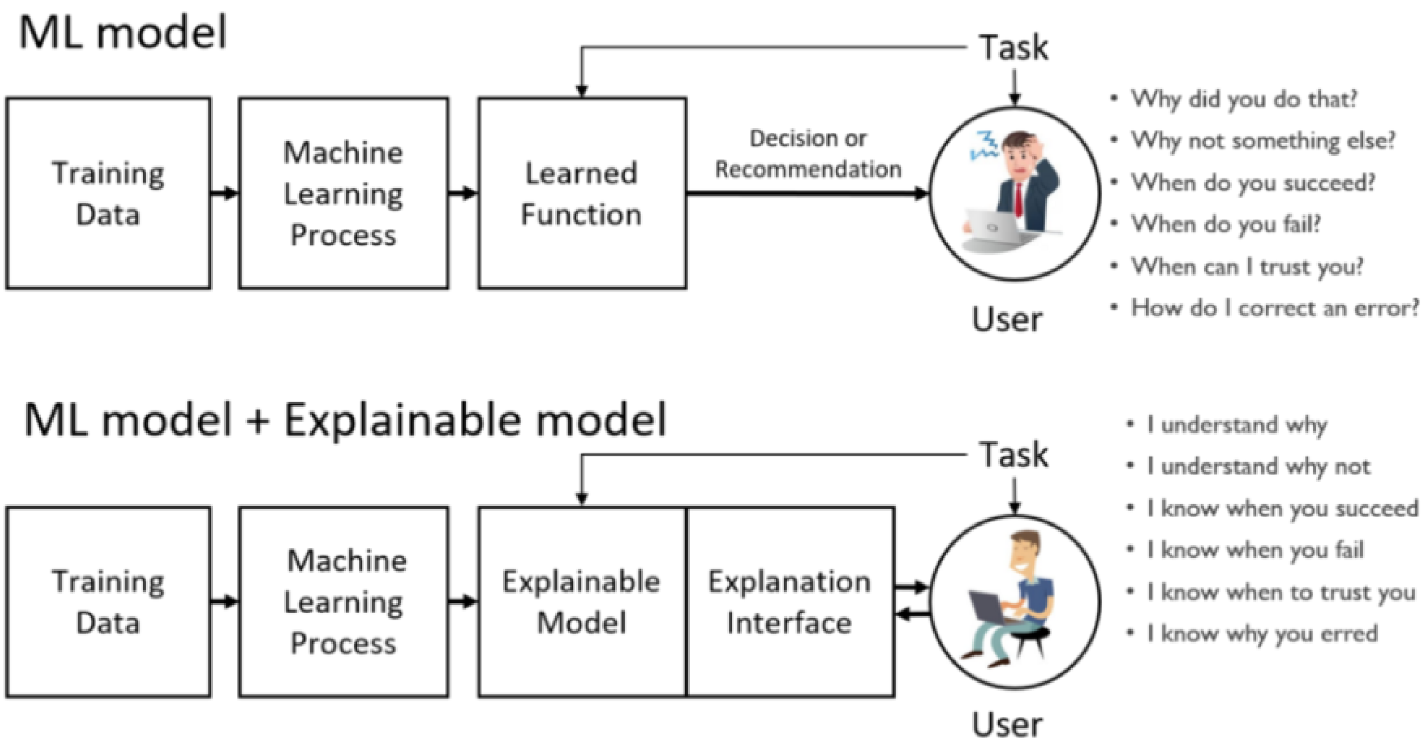
RESPONSIBLE AI: A STUDY ON GERMAN CREDIT CARD DATASET

By: J Krishna Ravali (19MBMB12)

Problem statement

As black-box Machine Learning (ML) models are increasingly being employed to make important predictions in critical contexts, the demand for transparency is increasing from the various stakeholders in AI



Source: Broad Agency Announcement Explainable Artificial Intelligence (XAI) DARPA-BAA-16–53

Fig.1. AI vs XAI

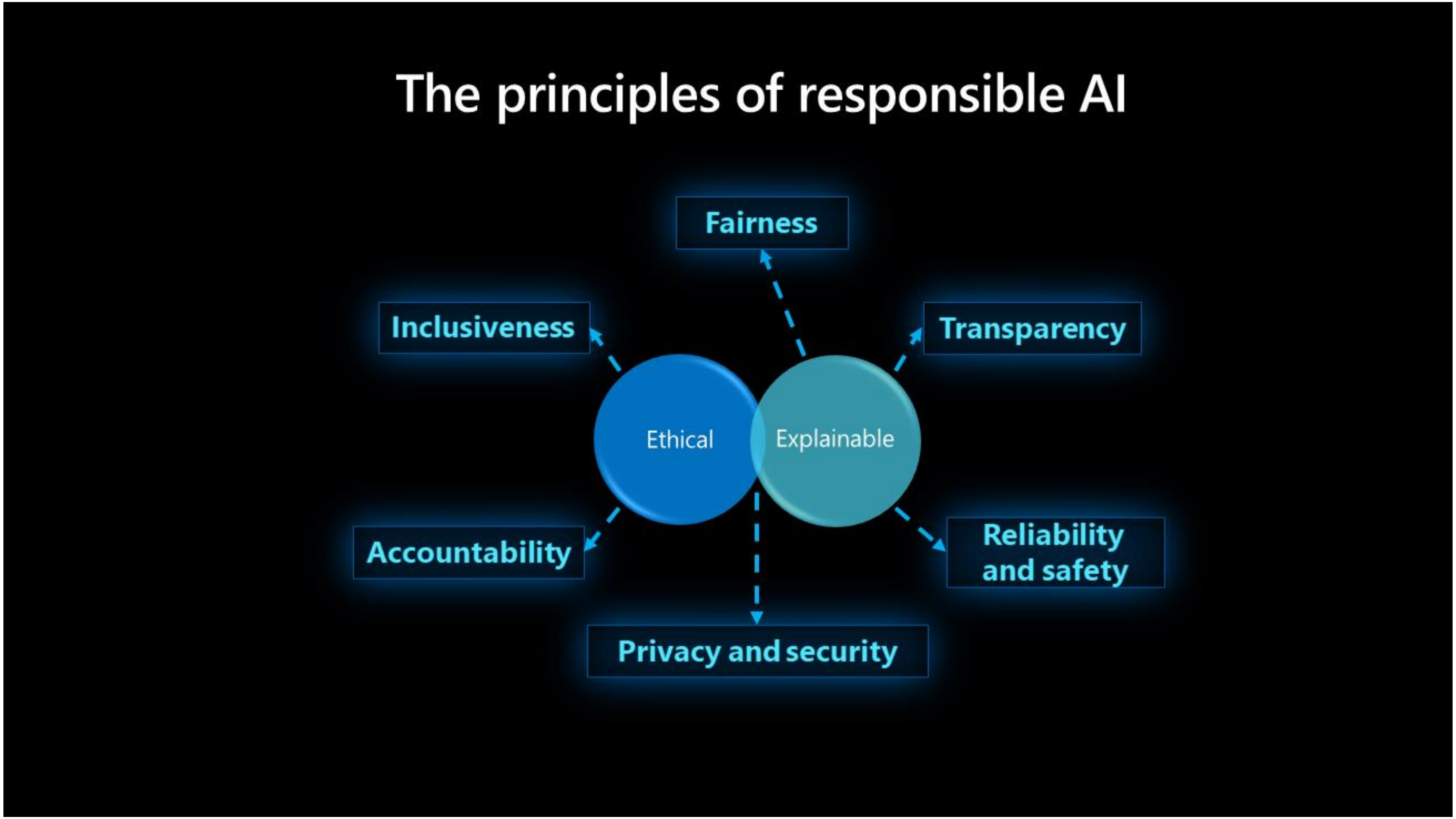
Motivation – Failure of AI

- Amazon's AI-Powered Recruiting Tool
- LG's IoT AI Assistant Cloi
- Microsoft's AI Chatbot Tay
- Tesla's autonomous cars failure in traffic
- Facial Recognition Failure In China

Reference: https://arxiv.org/pdf/2005.12379.pdf

Need for Responsible AI

- AI systems are designed to act autonomously in our world to make quicker and better decisions than humans
- Since the decisions derived from such systems ultimately affect human's lives (e.g. medicine, law or defense), there is an emerging need for understanding how such decisions are made by AI.
- Hence, we need to ***ensure that the purpose put into the machine is the purpose which we really want***.



Source: https://docs.microsoft.com/en-us/azure/cloud-adoption-framework/innovate/best-practices/trusted-ai

Data Description

german.data - Notepad

File Edit Format View Help

```
A11 6 A34 A43 1169 A65 A75 4 A93 A101 4 A121 67 A143 A152 2 A173 1 A192 A201 1
A12 48 A32 A43 5951 A61 A73 2 A92 A101 2 A121 22 A143 A152 1 A173 1 A191 A201 2
A14 12 A34 A46 2096 A61 A74 2 A93 A101 3 A121 49 A143 A152 1 A172 2 A191 A201 1
A11 42 A32 A42 7882 A61 A74 2 A93 A103 4 A122 45 A143 A153 1 A173 2 A191 A201 1
A11 24 A33 A40 4870 A61 A73 3 A93 A101 4 A124 53 A143 A153 2 A173 2 A191 A201 2
A14 36 A32 A46 9055 A65 A73 2 A93 A101 4 A124 35 A143 A153 1 A172 2 A192 A201 1
A14 24 A32 A42 2835 A63 A75 3 A93 A101 4 A122 53 A143 A152 1 A173 1 A191 A201 1
A12 36 A32 A41 6948 A61 A73 2 A93 A101 2 A123 35 A143 A151 1 A174 1 A192 A201 1
```

	CurrentAcc	NumMonths	CreditHistory	Purpose	CreditAmount	Savings	EmployDuration	PayBackPercent	Gender	Debtors	...	Collateral	Age	OtherPay
0	A11	6	A34	A43	1169	A65	A75	4	A93	A101	...	A121	67	
1	A12	48	A32	A43	5951	A61	A73	2	A92	A101	...	A121	22	
2	A14	12	A34	A46	2096	A61	A74	2	A93	A101	...	A121	49	
3	A11	42	A32	A42	7882	A61	A74	2	A93	A103	...	A122	45	
4	A11	24	A33	A40	4870	A61	A73	3	A93	A101	...	A124	53	

5 rows × 21 columns

- Source: UCI ML Repository
  https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29
- This dataset contains 1000 instances and 21 fields with both numerical and categorical field
- Target Field : CreditStatus
- Task: To predict if the credit score status of the user is good(1)/bad(0) using fair ML approach apart from adding explainability to the model outcome.

Field	Field Description
CurrentAcc	Status of checking existing account
NumMonths	Duration in months
CreditHistory	Credit History
Purpose	Purpose
CreditAmount	Credit Amount
Savings	Savings account
EmployDuration	Present employment
PayBackPercent	Installment rate in percentage of disposable income
Gender	Personal status and sex
Debtors	Other debtors/ guarantors
ResidenceDuration	Present residence since
Collateral	Collateral property
Age	Age in years
OtherPayBackPlan	Other installments plan
Property	Housing
ExistingCredit	Existing credit at this bank
Job	Job
Dependents	Number of people being liable to provide maintenance for
Telephone	Telephone
Foreignworker	Foreign worker
CreditStatus	Credit Status (Target field: good/bad)

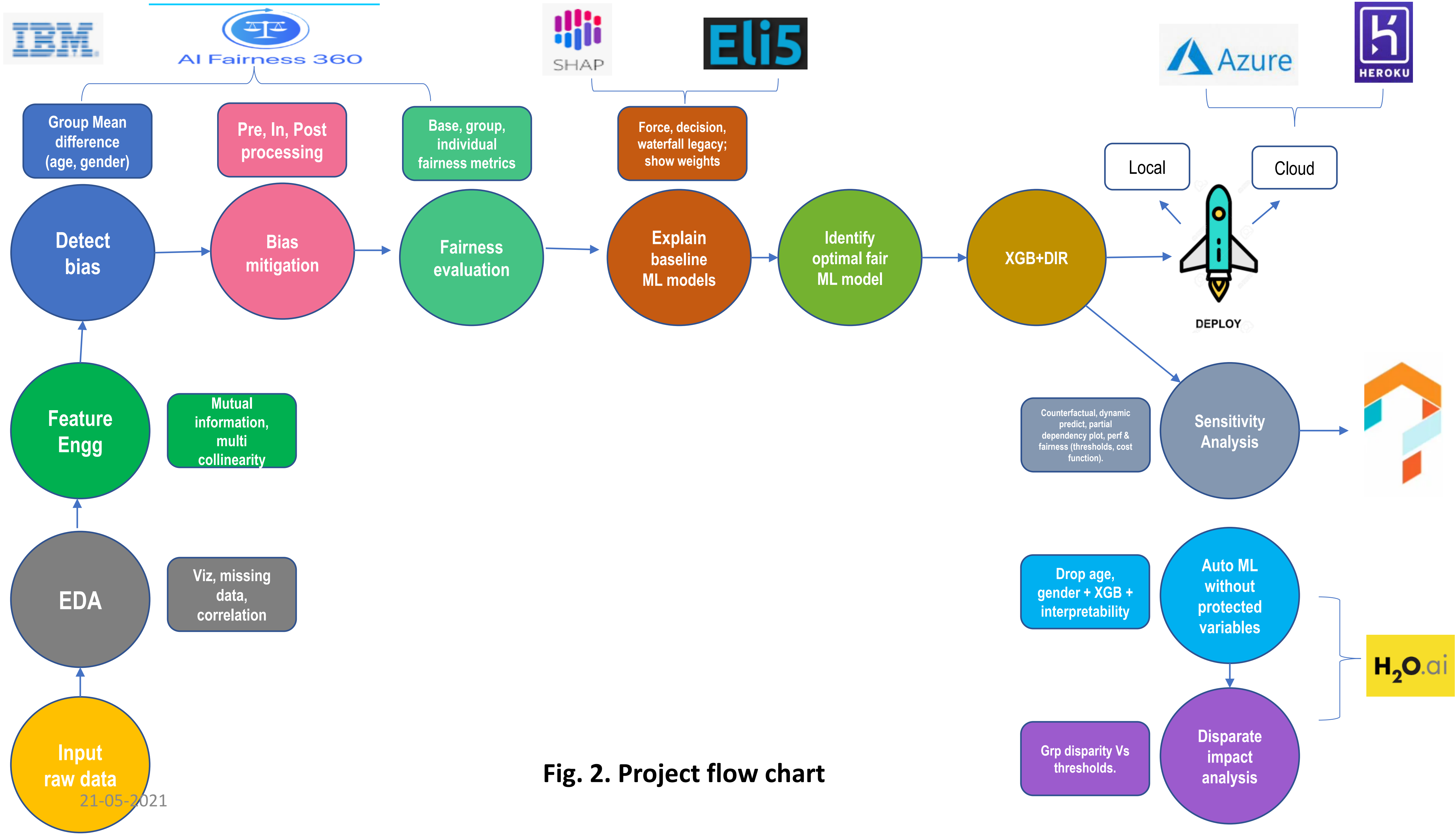


Fig. 2. Project flow chart

21-05-2021

RESULTS

eXplainable AI (XAI) (w.r.to age)

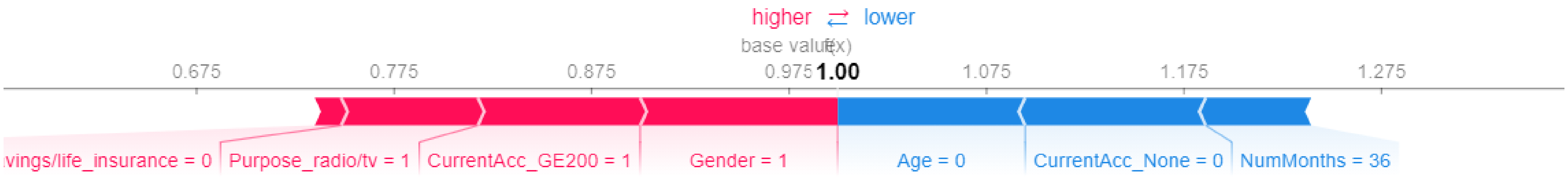
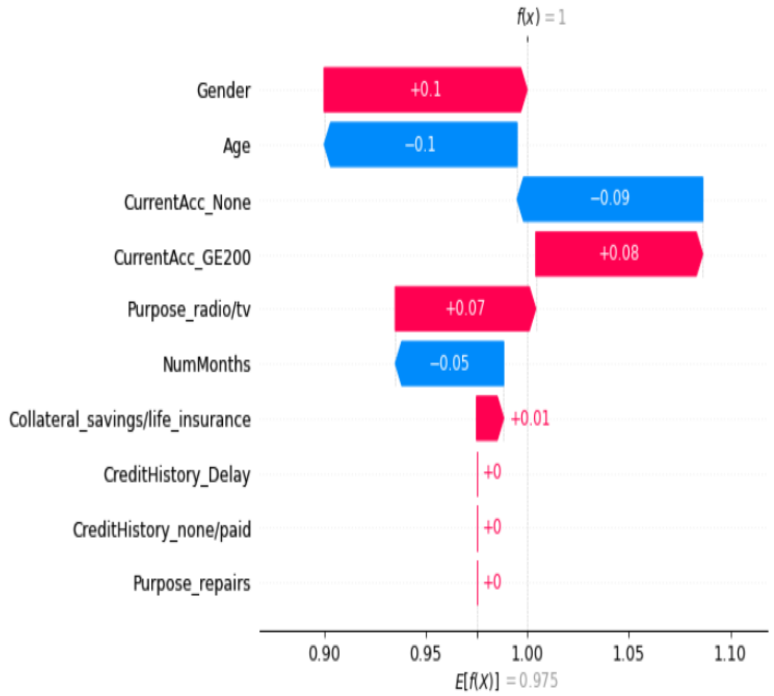


Fig. 3.a. Shap- Force plot



Weight	Feature
0.0263 ± 0.0198	NumMonths
0.0146 ± 0.0131	Purpose_radio/tv
0.0102 ± 0.0117	Age
0.0073 ± 0.0092	Gender
0.0073 ± 0.0000	CurrentAcc_GE200
0.0029 ± 0.0072	CreditHistory_none/paid
0.0015 ± 0.0058	CurrentAcc_None
0 ± 0.0000	Purpose_repairs
0 ± 0.0000	Collateral_savings/life_insurance
0 ± 0.0000	CreditHistory_Delay

Fig. 4. ELI5 – Feature importance plot (show_weights)

Fig. 3.b. Shap- waterfall legacy plot

Fairness metrics w.r.to age (before and after bias mitigation)

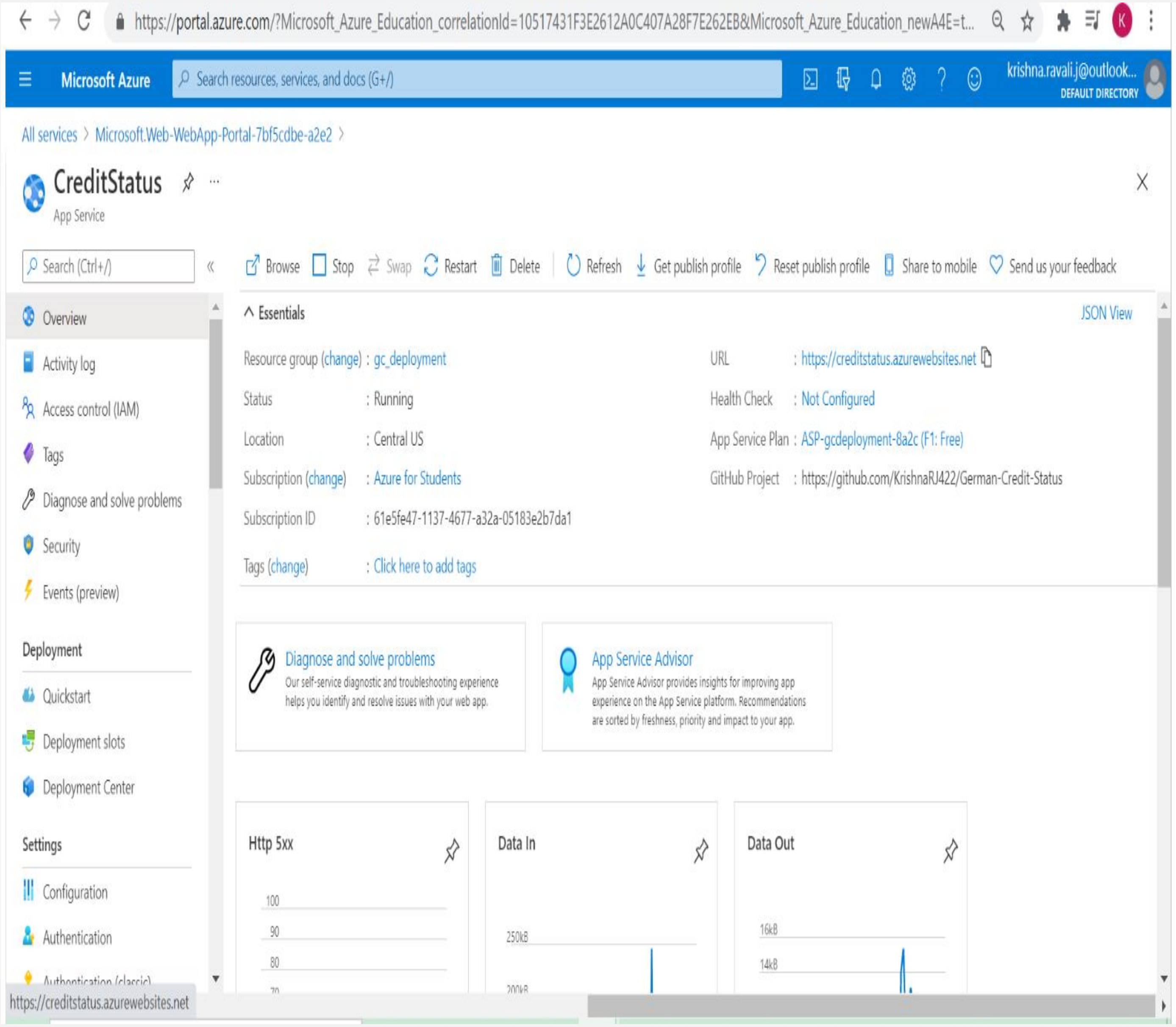
SL.NO	MODEL	BIAS MITIGATION TECH/FAIRNESS METRIC	Accuracy	F1	DI	SPD	EOD	AOD	ERD	CNT	TI
SL.NO	model names	**objective**	1	1	1	0	0	0	0	1	0
GC1	RANDOM FOREST WITH HYPER PARAMS	MODEL VALUE	0.7	0.819277	0.876152	-0.12312	-0.04762	-0.16438	-0.04963	0.975	0.062379
GC1	RANDOM FOREST WITH HYPER PARAMS	REWEIGHTING	0.680576	0.806363	0.907166	-0.09222	-0.04762	-0.16438	-0.16635	0.975	0.062379
GC1	RANDOM FOREST WITH HYPER PARAMS	DIR	0.675	0.8	0.612903	-0.3871	-0.33333	-0.41667	0.073487	0.946	0.093203
GC1	RANDOM FOREST WITH HYPER PARAMS	AD	0.68	0.808383	0.903226	-0.09677	-0.09524	-0.09762	0.041229	0.98	0.068214
GC1	RANDOM FOREST WITH HYPER PARAMS	PRR	0.685	0.796117	0.55914	-0.40695	-0.44171	-0.38595	0.199847	0.893	0.127789
GC1	RANDOM FOREST WITH HYPER PARAMS	EO	0.485	0.565401	0.962049	-0.01909	-0.07143	0.009569	0.077687	0.602	0.485228
GC1	RANDOM FOREST WITH HYPER PARAMS	CEO	0.675	0.8	0.612903	-0.3871	-0.33333	-0.41667	0.073487	0.946	0.093203
GC1	RANDOM FOREST WITH HYPER PARAMS	ROC	0.69	0.75969	1.015497	0.009353	-0.05747	0.051453	0.091239	0.776	0.256668
GC2	XGBOOST WITH HYPER PARAMS	MODEL VALUE	0.73	0.83125	0.943548	-0.05211	-0.078	-0.03522	0.062226	0.963	0.074753
GC2	XGBOOST WITH HYPER PARAMS	REWEIGHTING	0.717561	0.821422	0.964971	-0.03207	-0.078	-0.03522	-0.05262	0.963	0.074753
GC2	XGBOOST WITH HYPER PARAMS	DIR	0.73	0.833333	0.91996	-0.07578	-0.039	-0.09403	-0.01412	0.961	0.065206
GC2	XGBOOST WITH HYPER PARAMS	AD	0.685	0.813056	1	0	0	0	0.008971	1	0.058241
GC2	XGBOOST WITH HYPER PARAMS	PRR	0.685	0.796117	0.55914	-0.40695	-0.44171	-0.38595	0.199847	0.893	0.127789
GC2	XGBOOST WITH HYPER PARAMS	EO	0.735	0.835913	0.966109	-0.03169	0.017241	-0.05647	-0.04638	0.936	0.064804
GC2	XGBOOST WITH HYPER PARAMS	CEO	0.73	0.833333	0.91996	-0.07578	-0.039	-0.09403	-0.01412	0.961	0.065206
GC2	XGBOOST WITH HYPER PARAMS	ROC	0.665	0.730924	0.975552	-0.01374	-0.10961	0.044254	0.138003	0.837	0.299887
GC3	XGBOOST	MODEL VALUE	0.67	0.769231	0.702102	-0.23268	-0.21716	-0.23877	0.06757	0.809	0.196757
GC3	XGBOOST	REWEIGHTING	0.653423	0.754711	0.742212	-0.19956	-0.21716	-0.23877	0.045669	0.809	0.196757
GC3	XGBOOST	DIR	0.705	0.802676	0.681452	-0.27143	-0.24713	-0.28205	0.070815	0.845	0.140992
GC3	XGBOOST	AD	0.685	0.813056	1	0	0	0	0.008971	1	0.058241
GC3	XGBOOST	PRR	0.685	0.796117	0.55914	-0.40695	-0.44171	-0.38595	0.199847	0.893	0.127789
GC3	XGBOOST	EO	0.68	0.774648	0.959484	-0.02997	0.064039	-0.0793	-0.11147	0.74	0.195462
GC3	XGBOOST	CEO	0.68	0.802469	0.580645	-0.41935	-0.33333	-0.46667	0.041229	0.945	0.092932
GC3	XGBOOST	ROC	0.65	0.700855	1.076862	0.036839	0.02422	0.048902	0.005726	0.756	0.353109
GC4	RANDOM FOREST	MODEL VALUE	0.675	0.779661	0.745185	-0.20958	-0.09154	-0.27313	-0.07921	0.817	0.169886
GC4	RANDOM FOREST	REWEIGHTING	0.656674	0.765708	0.816155	-0.15049	-0.09154	-0.27313	-0.11065	0.817	0.169886
GC4	RANDOM FOREST	DIR	0.715	0.80678	0.745185	-0.20958	-0.18227	-0.22132	0.044474	0.8	0.144795
GC4	RANDOM FOREST	AD	0.685	0.813056	1	0	0	0	0.008971	1	0.058241
GC4	RANDOM FOREST	PRR	0.685	0.796117	0.55914	-0.40695	-0.44171	-0.38595	0.199847	0.893	0.127789
GC4	RANDOM FOREST	EO	0.705	0.792982	0.951869	-0.03588	0.038177	-0.07336	-0.08189	0.736	0.176413
GC4	RANDOM FOREST	CEO	0.685	0.806154	0.612903	-0.3871	-0.28571	-0.44286	0.008971	0.943	0.08768
GC4	RANDOM FOREST	ROC	0.7	0.777778	1.022177	0.014697	0.050903	-0.00096	-0.04963	0.755	0.218245
GC5	KNN	MODEL VALUE	0.665	0.761566	0.681452	-0.24127	-0.1913	-0.2664	0.023478	0.79	0.21339
GC5	KNN	REWEIGHTING	0.648295	0.746982	0.738285	-0.19648	-0.1913	-0.2664	0.015292	0.79	0.21339
GC5	KNN	DIR	0.68	0.766423	0.67028	-0.23802	-0.2303	-0.23873	0.079404	0.837	0.221807
GC5	KNN	AD	0.685	0.813056	1	0	0	0	0.008971	1	0.058241
GC5	KNN	PRR	0.685	0.796117	0.55914	-0.40695	-0.44171	-0.38595	0.199847	0.893	0.127789
GC5	KNN	EO	0.635	0.715953	0.962049	-0.02291	-0.00575	-0.02929	-0.01203	0.739	0.300942
GC5	KNN	CEO	0.675	0.798762	0.548387	-0.45161	-0.38095	-0.49048	0.073487	0.941	0.098209
GC5	KNN	ROC	0.695	0.781362	0.999462	-0.00038	0.01642	-0.00594	-0.01737	0.83	0.198467
GC6	LOGISTIC REGRESSION	MODEL VALUE	0.72	0.816993	0.60972	-0.35102	-0.34647	-0.35059	0.126742	0.89	0.114498
GC6	LOGISTIC REGRESSION	REWEIGHTING	0.695573	0.79851	0.644791	-0.31652	-0.34647	-0.35059	0.096001	0.89	0.114498
GC6	LOGISTIC REGRESSION	DIR	0.72	0.816993	0.60972	-0.35102	-0.34647	-0.35059	0.126742	0.89	0.114498
GC6	LOGISTIC REGRESSION	AD	0.645	0.76412	1.171087	0.136667	0.077176	0.170663	0.037984	0.829	0.172612
GC6	LOGISTIC REGRESSION	PRR	0.685	0.796117	0.55914	-0.40695	-0.44171	-0.38595	0.199847	0.893	0.127789
GC6	LOGISTIC REGRESSION	EO	0.63	0.727941	0.948107	-0.03531	-0.06609	-0.01606	0.058408	0.696	0.262104
GC6	LOGISTIC REGRESSION	CEO	0.675	0.798762	0.548387	-0.45161	-0.38095	-0.49048	0.073487	0.943	0.098209
GC6	LOGISTIC REGRESSION	ROC	0.695	0.769811	1.069943	0.044283	0.020525	0.06215	0.020806	0.793	0.234754
GC7	SVM	MODEL VALUE	0.725	0.826498	0.879292	-0.11071	-0.117	-0.10472	0.056308	0.919	0.084797
GC7	SVM	REWEIGHTING	0.710024	0.815003	0.906863	-0.08475	-0.117	-0.10472	-0.04055	0.919	0.084797
GC7	SVM	DIR	0.725	0.826498	0.879292	-0.11071	-0.117	-0.10472	0.056308	0.919	0.084797
GC7	SVM	AD	0.645	0.76412	1.171087	0.136667	0.077176	0.170663	0.037984	0.829	0.172612
GC7	SVM	PRR	0.685	0.796117	0.55914	-0.40695	-0.44171	-0.38595	0.199847	0.893	0.127789
GC7	SVM	EO	0.675	0.782609	0.99482	-0.0042	0.003695	-0.00664	-0.00286	0.811	0.159476
GC7	SVM	CEO	0.685	0.809668	0.806452	-0.19355	-0.14286	-0.22143	0.008971	0.966	0.07289
GC7	SVM	ROC	0.64	0.707317	0.937912	-0.03417	-0.07512	-0.00737	0.070242	0.825	0.32814

Fairness metrics w.r.to gender (before and after bias mitigation)

SL.NO	MODEL	BIAS MITIGATION TECH/FAIRNESS METRIC	Accuracy	F1	DI	SPD	EOD	AOD	ERD	CNT	TI
SL.NO	model names	**objective**	**1**	**1**	**1**	**0**	**0**	**0**	**0**	**1**	**0**
GC1	RANDOM FOREST WITH HYPER PARAMS	MODEL VALUE	0.7	0.819277	0.957562	-0.04186	-0.02439	-0.06912	-0.10654	0.975	0.062379
GC1	RANDOM FOREST WITH HYPER PARAMS	REWEIGHTING	0.684158	0.808162	0.966952	-0.03254	-0.02439	-0.06912	-0.1959	0.975	0.062379
GC1	RANDOM FOREST WITH HYPER PARAMS	DIR	0.705	0.821752	0.964258	-0.03501	-0.02439	-0.05912	-0.0997	0.973	0.062101
GC1	RANDOM FOREST WITH HYPER PARAMS	AD	0.69	0.813253	0.907407	-0.09259	-0.04878	-0.13977	-0.12024	0.977	0.067752
GC1	RANDOM FOREST WITH HYPER PARAMS	PRR	0.71	0.811688	0.798822	-0.18189	-0.15346	-0.24596	-0.04211	0.893	0.115516
GC1	RANDOM FOREST WITH HYPER PARAMS	EO	0.695	0.81571	0.990089	-0.00964	-0.01397	-0.01545	-0.08803	0.975	0.067496
GC1	RANDOM FOREST WITH HYPER PARAMS	CEO	0.69	0.814371	0.944444	-0.05556	-0.02439	-0.08912	-0.12024	0.985	0.062882
GC1	RANDOM FOREST WITH HYPER PARAMS	ROC	0.67	0.725	0.960526	-0.02055	-0.07088	-0.05006	0.055302	0.84	0.319948
GC2	XGBOOST WITH HYPER PARAMS	MODEL VALUE	0.73	0.83125	1.01643	0.014967	0.00686	-0.01195	-0.09082	0.963	0.074753
GC2	XGBOOST WITH HYPER PARAMS	REWEIGHTING	0.71875	0.82244	1.037968	0.034313	0.00686	-0.01195	-0.16747	0.963	0.074753
GC2	XGBOOST WITH HYPER PARAMS	DIR	0.73	0.83125	0.961317	-0.03577	-0.02795	-0.07782	-0.09082	0.942	0.074753
GC2	XGBOOST WITH HYPER PARAMS	AD	0.685	0.813056	1	0	0	0	-0.10173	1	0.058241
GC2	XGBOOST WITH HYPER PARAMS	PRR	0.71	0.811688	0.798822	-0.18189	-0.15346	-0.24596	-0.04211	0.893	0.115516
GC2	XGBOOST WITH HYPER PARAMS	EO	0.715	0.82243	1.008845	0.008118	-0.01753	0.004312	-0.06063	0.935	0.08078
GC2	XGBOOST WITH HYPER PARAMS	CEO	0.71	0.823171	0.907537	-0.09056	-0.04878	-0.14824	-0.11821	0.954	0.066623
GC2	XGBOOST WITH HYPER PARAMS	ROC	0.675	0.728033	1.074074	0.037037	-0.03608	0.025808	0.062151	0.876	0.318399
GC3	XGBOOST	MODEL VALUE	0.67	0.769231	0.769157	-0.18341	-0.20605	-0.20225	0.055302	0.809	0.196757
GC3	XGBOOST	REWEIGHTING	0.655381	0.755867	0.792456	-0.16308	-0.20605	-0.20225	0.025807	0.809	0.196757
GC3	XGBOOST	DIR	0.685	0.790698	0.707123	-0.26078	-0.23018	-0.32278	-0.00025	0.853	0.148183
GC3	XGBOOST	AD	0.685	0.813056	1	0	0	0	-0.10173	1	0.058241
GC3	XGBOOST	PRR	0.71	0.811688	0.798822	-0.18189	-0.15346	-0.24596	-0.04211	0.893	0.115516
GC3	XGBOOST	EO	0.665	0.780328	0.98916	-0.00913	-0.02134	-0.01606	-0.05302	0.824	0.149962
GC3	XGBOOST	CEO	0.67	0.791139	0.633972	-0.36352	-0.29268	-0.44403	-0.0208	0.908	0.118718
GC3	XGBOOST	ROC	0.67	0.736	1.06813	0.037798	0.051067	-0.03908	-0.07154	0.798	0.293243
GC4	RANDOM FOREST	MODEL VALUE	0.675	0.779661	0.856173	-0.11821	-0.0841	-0.18128	-0.06469	0.817	0.169886
GC4	RANDOM FOREST	REWEIGHTING	0.660548	0.767725	0.884713	-0.09414	-0.0841	-0.18128	-0.10477	0.817	0.169886
GC4	RANDOM FOREST	DIR	0.665	0.778878	0.802662	-0.17301	-0.08054	-0.27796	-0.12912	0.846	0.155172
GC4	RANDOM FOREST	AD	0.685	0.813056	1	0	0	0	-0.10173	1	0.058241
GC4	RANDOM FOREST	PRR	0.71	0.811688	0.798822	-0.18189	-0.15346	-0.24596	-0.04211	0.893	0.115516
GC4	RANDOM FOREST	EO	0.645	0.773163	1.013889	0.012177	-0.00737	0.027855	-0.05505	0.875	0.140876
GC4	RANDOM FOREST	CEO	0.685	0.802508	0.713477	-0.2826	-0.18471	-0.39004	-0.10173	0.917	0.10261
GC4	RANDOM FOREST	ROC	0.635	0.672646	1.108075	0.045155	-0.0155	0.027636	0.058092	0.78	0.396715
GC5	KNN	MODEL VALUE	0.665	0.761566	0.772487	-0.17453	-0.1748	-0.21509	0.023085	0.79	0.21339
GC5	KNN	REWEIGHTING	0.650903	0.748534	0.802117	-0.15016	-0.1748	-0.21509	0.000303	0.79	0.21339
GC5	KNN	DIR	0.62	0.739726	0.596092	-0.35134	-0.32444	-0.38991	0.062912	0.85	0.212702
GC5	KNN	AD	0.685	0.813056	1	0	0	0	-0.10173	1	0.058241
GC5	KNN	PRR	0.71	0.811688	0.798822	-0.18189	-0.15346	-0.24596	-0.04211	0.893	0.115516
GC5	KNN	EO	0.675	0.790997	1.000583	0.000507	0.041413	-0.05314	-0.14079	0.87	0.128574
GC5	KNN	CEO	0.645	0.776025	0.62963	-0.37037	-0.34146	-0.4015	0.046423	0.919	0.130436
GC5	KNN	ROC	0.55	0.575472	1.05144	0.019026	-0.00889	0.000938	0.043125	0.778	0.511702
GC6	LOGISTIC REGRESSION	MODEL VALUE	0.72	0.816993	0.757856	-0.21892	-0.18826	-0.29182	-0.02841	0.89	0.114498
GC6	LOGISTIC REGRESSION	REWEIGHTING	0.699632	0.801753	0.790647	-0.18759	-0.18826	-0.29182	-0.07014	0.89	0.114498
GC6	LOGISTIC REGRESSION	DIR	0.72	0.816993	0.757856	-0.21892	-0.18826	-0.29182	-0.02841	0.89	0.114498
GC6	LOGISTIC REGRESSION	AD	0.705	0.820669	0.952008	-0.04668	-0.01397	-0.09237	-0.12506	0.968	0.066931
GC6	LOGISTIC REGRESSION	PRR	0.71	0.811688	0.798822	-0.18189	-0.15346	-0.24596	-0.04211	0.893	0.115516
GC6	LOGISTIC REGRESSION	EO	0.69	0.797386	1.011141	0.009386	-0.01778	0.005723	-0.04414	0.823	0.132424
GC6	LOGISTIC REGRESSION	CEO	0.695	0.806349	0.709483	-0.28057	-0.2091	-0.37224	-0.06266	0.902	0.106886
GC6	LOGISTIC REGRESSION	ROC	0.655	0.701299	1.033769	0.015728	-0.04319	-0.00621	0.060122	0.857	0.35686
GC7	SVM	MODEL VALUE	0.725	0.826498	0.848608	-0.14206	-0.11153	-0.20654	-0.0723	0.919	0.084797
GC7	SVM	REWEIGHTING	0.706444	0.813123	0.875846	-0.11571	-0.11153	-0.20654	-0.13226	0.919	0.084797
GC7	SVM	DIR	0.72	0.822785	0.854847	-0.13521	-0.10112	-0.20133	-0.07915	0.914	0.090076
GC7	SVM	AD	0.705	0.820669	0.952008	-0.04668	-0.01397	-0.09237	-0.12506	0.968	0.066931
GC7	SVM	PRR	0.71	0.811688	0.798822	-0.18189	-0.15346	-0.24596	-0.04211	0.893	0.115516
GC7	SVM	EO	0.715	0.821317	0.996101	-0.00355	-0.00711	-0.02894	-0.086	0.91	0.085635
GC7	SVM	CEO	0.7	0.8125	0.805359	-0.18798	-0.13592	-0.25873	-0.08118	0.914	0.091659
GC7	SVM	ROC	0.69	0.752	1.06813	0.037798	-0.00457	0.00156	0.006596	0.847	0.27749

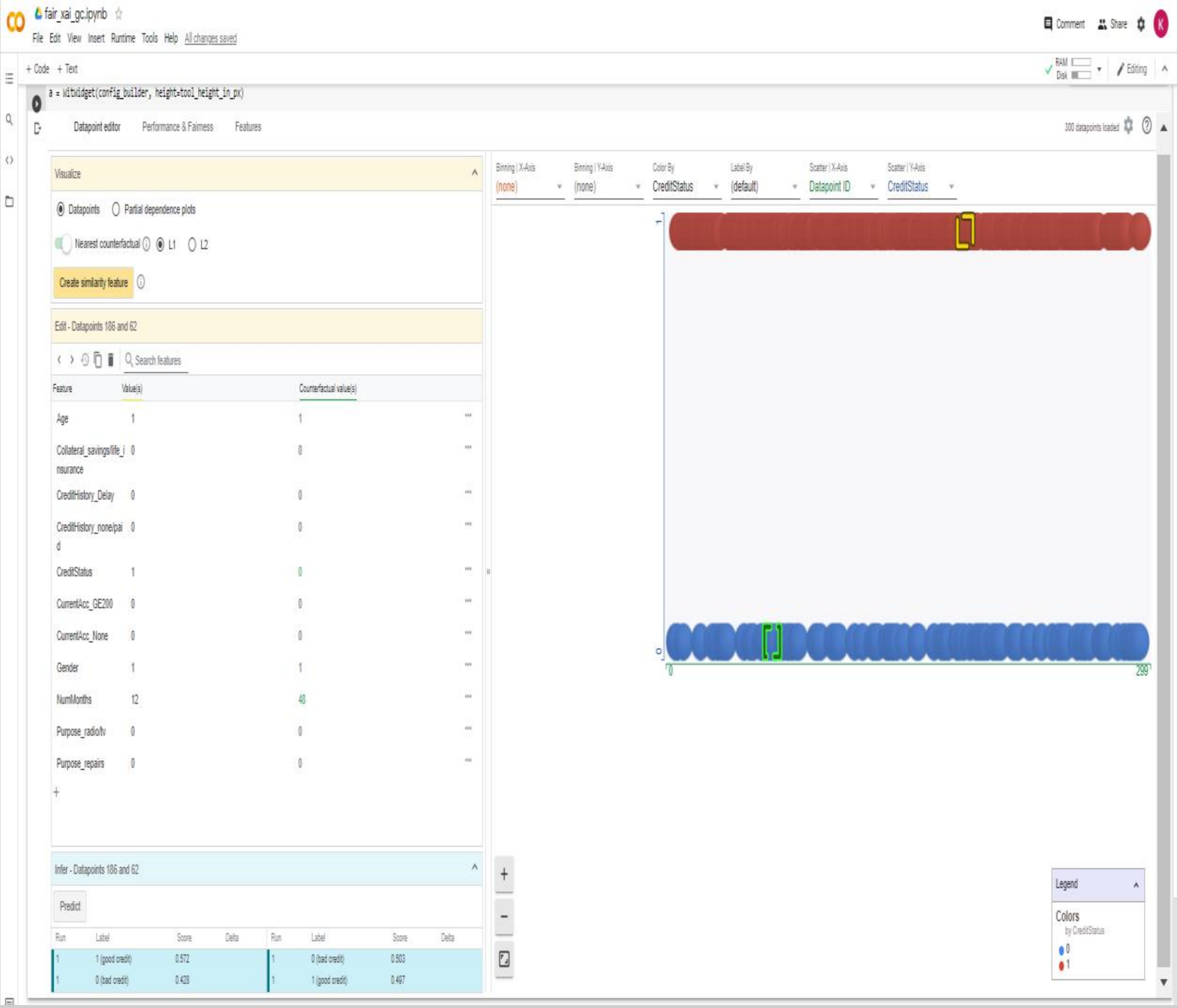
Deployment

- The XBG+DIR model is deployed using FLASK:
  - In local (127.0.0.1:5000)
  - In cloud (as REST API):
    - In Heroku (https://credit-score-status.herokuapp.com/)
    - In Azure (https://creditstatus.azurewebsites.net/)

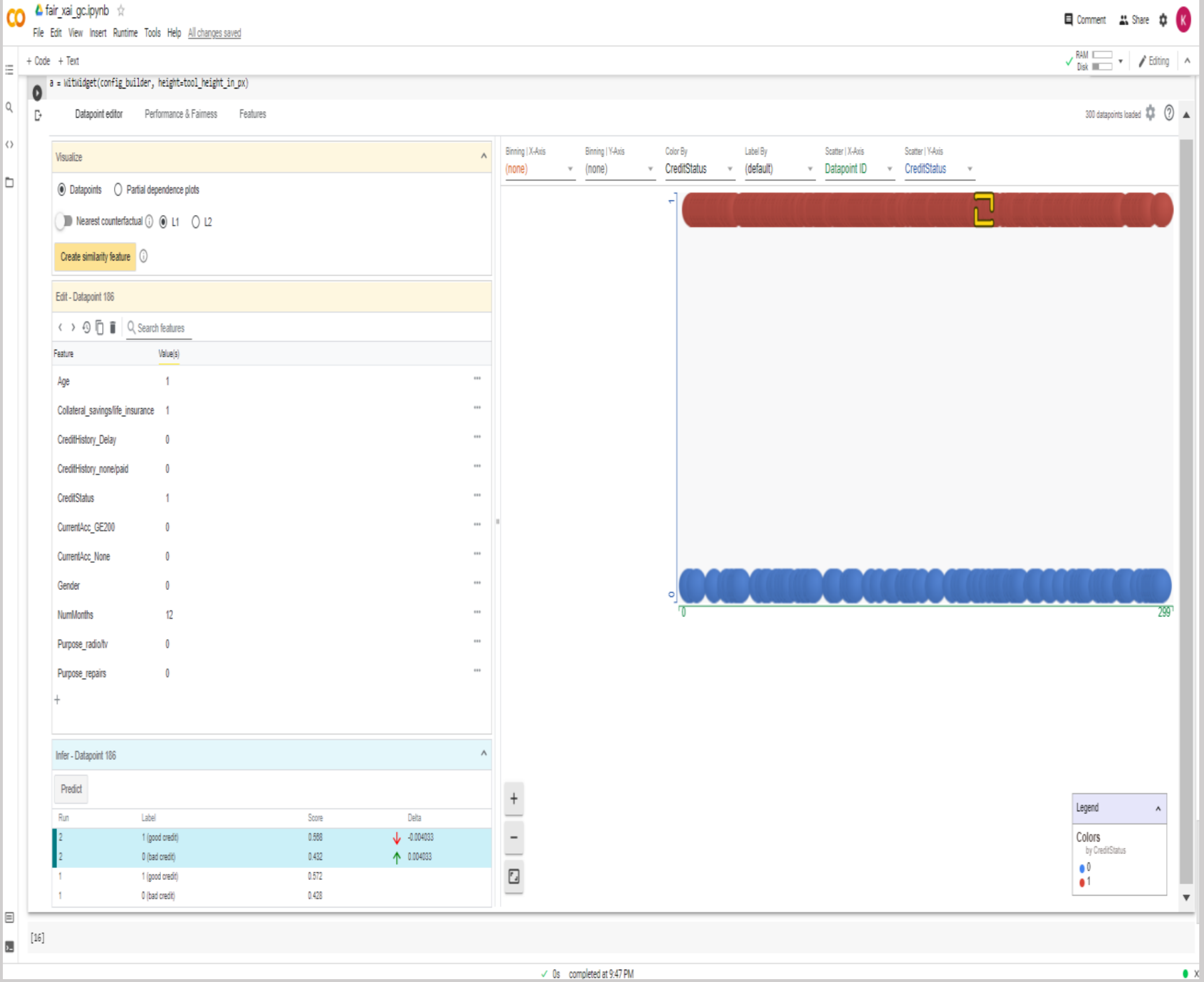


Heroku cloud application

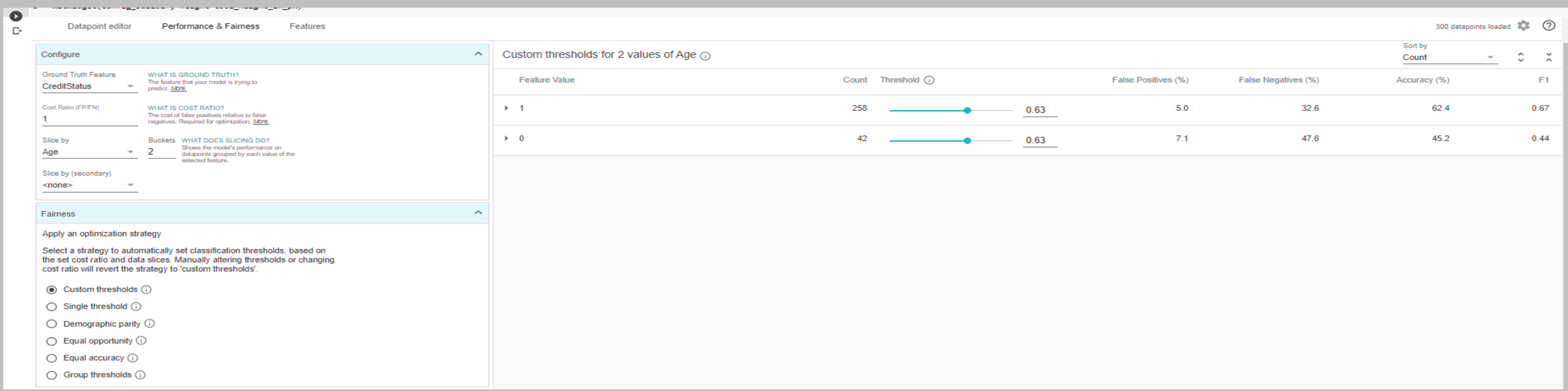
Sensitivity Analysis (Using Google's What-if tool)



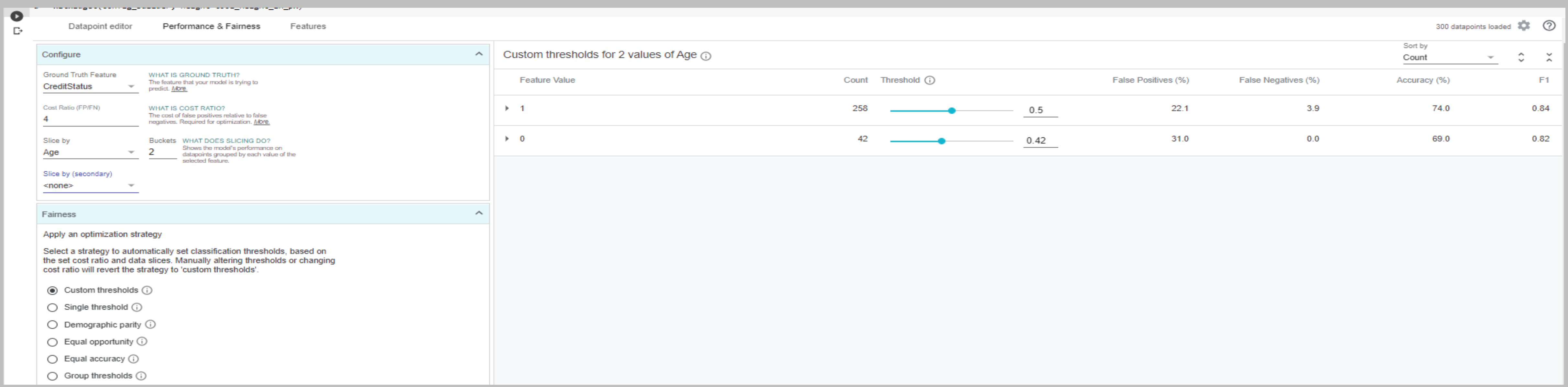
Counterfactual data points are the neighbour data points that got different classification. For the data point in yellow and green in the right pane, the only attribute that is different is NumMonths (Yellow data point has Nummonths=12 and green data point has Nummonths=48) whose value is highlighted in green in left pane has got different classification which is given at the bottom, yellow data point has got higher probability for good credit and green data point has higher probability for bad credit.



WIT allows us to dynamically alter/toggle a feature value and check the model prediction. In the above figure the gender and collateral values of data point 186 are altered and prediction of the model is observed where the score of outcomes for good credit and bad credit has changed from 0.572 to 0.568 and 0.428 to 0.432 respectively with this change.



Model set threshold and cost function.



Improved performance with altered cost function (as FP 4x costlier than FN) and threshold values

Disparate Impact Analysis (using H2O.ai tool)

Fig. 6. view at the initiation of Auto ML

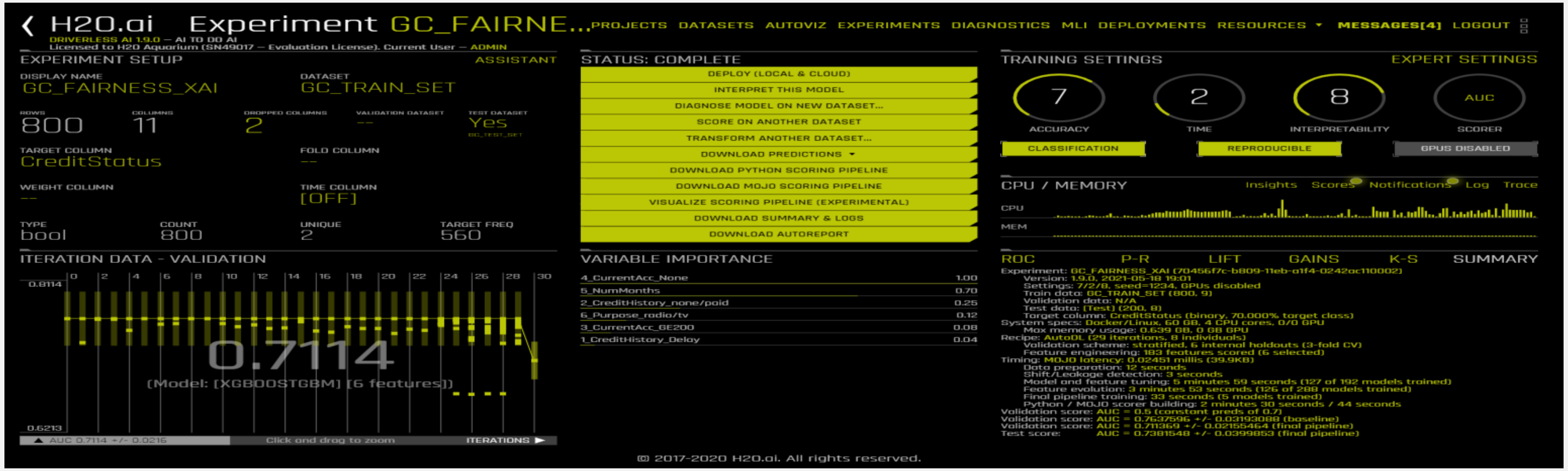
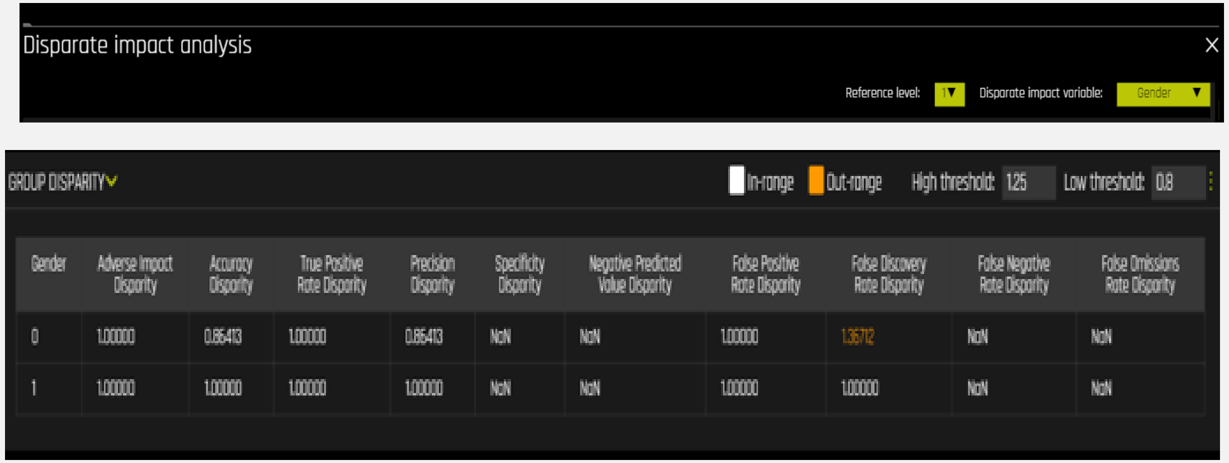
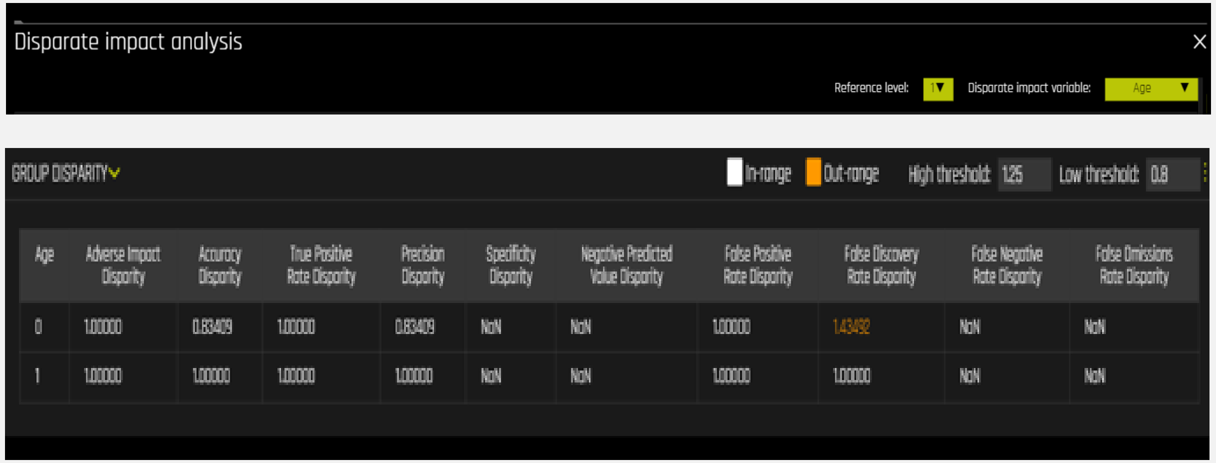


Fig. 7. view after completion



Gender	Adverse Impact Disparity	Accuracy Disparity	True Positive Rate Disparity	Precision Disparity	Specificity Disparity	Negative Predicted Value Disparity	False Positive Rate Disparity	False Discovery Rate Disparity	False Negative Rate Disparity	False Omissions Rate Disparity
0	1.00000	0.86413	1.00000	0.86413	NaN	NaN	1.00000	1.36712	NaN	NaN
1	1.00000	1.00000	1.00000	1.00000	NaN	NaN	1.00000	1.00000	NaN	NaN

Fig. 8. Fairness metrics when threshold is set to standard default values (gender)



Age	Adverse Impact Disparity	Accuracy Disparity	True Positive Rate Disparity	Precision Disparity	Specificity Disparity	Negative Predicted Value Disparity	False Positive Rate Disparity	False Discovery Rate Disparity	False Negative Rate Disparity	False Omissions Rate Disparity
0	1.00000	0.83409	1.00000	0.83409	NaN	NaN	1.00000	1.43492	NaN	NaN
1	1.00000	1.00000	1.00000	1.00000	NaN	NaN	1.00000	1.00000	NaN	NaN

Fig. 9. Fairness metrics when threshold is set to standard default values (age)

Gender	Adverse Impact Disparity	Accuracy Disparity	True Positive Rate Disparity	Precision Disparity	Specificity Disparity	Negative Predicted Value Disparity	False Positive Rate Disparity	False Discovery Rate Disparity	False Negative Rate Disparity	False Omissions Rate Disparity
0	1.00000	0.86413	1.00000	0.86413	NaN	NaN	1.00000	1.36712	NaN	NaN
1	1.00000	1.00000	1.00000	1.00000	NaN	NaN	1.00000	1.00000	NaN	NaN

Fig. 10. Fairness Metrics when high threshold is set to 1.4 (gender)

At default set values, 1.25 to 0.8, group disparity metrics at gender=0 is highlighted and begin to flag, which implies Female will be treated unfairly by the model.

We can see that, if we adjust the high threshold to 1.4 from 1.25, the group disparity metrics looks fine which implies higher benefit is observed for disadvantageous group, female.

Age	Adverse Impact Disparity	Accuracy Disparity	True Positive Rate Disparity	Precision Disparity	Specificity Disparity	Negative Predicted Value Disparity	False Positive Rate Disparity	False Discovery Rate Disparity	False Negative Rate Disparity	False Omissions Rate Disparity
0	1.00000	0.83409	1.00000	0.83409	NaN	NaN	1.00000	1.43492	NaN	NaN
1	1.00000	1.00000	1.00000	1.00000	NaN	NaN	1.00000	1.00000	NaN	NaN

Fig. 11. Fairness Metrics when high threshold is set to 1.5 (age)

At default set values, 1.25 to 0.8, group disparity metrics at age=0 is highlighted and begin to flag, which implies age <26 will be treated unfairly by the model.

We can see that, if we adjust the high threshold to 1.5 from 1.25, the group disparity metrics looks fine which implies higher benefit is observed for disadvantageous group, age<26.

Recommendations & Future Scope

Recommendations:

- From the correlation plot w.r.to target field, it has been observed that, there aren't fields of higher correlation with target field which might have led to lower model accuracy. So, adding fields that could explain variance in target variables better to the dataset can improve model predictive power
- Also, the dataset used to train this model is very small, so adding more historical data points can improve model performance.

Future Scope:

- Replication of deployment process for protected attribute gender
- A docker container and image is created for fair model, which need to be deployed to AWS
- Exploring other facets of Responsible AI.
- Integrating WIT with google cloud deployed fair ML model.

References

- https://ethical.institute/principles.html
- https://analyticsindiamag.com/top-8-funniest-and-shocking-ai-failures-of-all-time/
- https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6
- https://shap.readthedocs.io/
- https://arxiv.org/pdf/2005.12379.pdf
- https://arxiv.org/pdf/1811.11154.pdf
- https://towardsdatascience.com/
- https://medium.com

Appendix (Github code links for this project)

- https://github.com/KrishnaRJ422/Explainability_Bias_Fairness-in-AI

- https://github.com/KrishnaRJ422/German-Credit-Status

THANK YOU

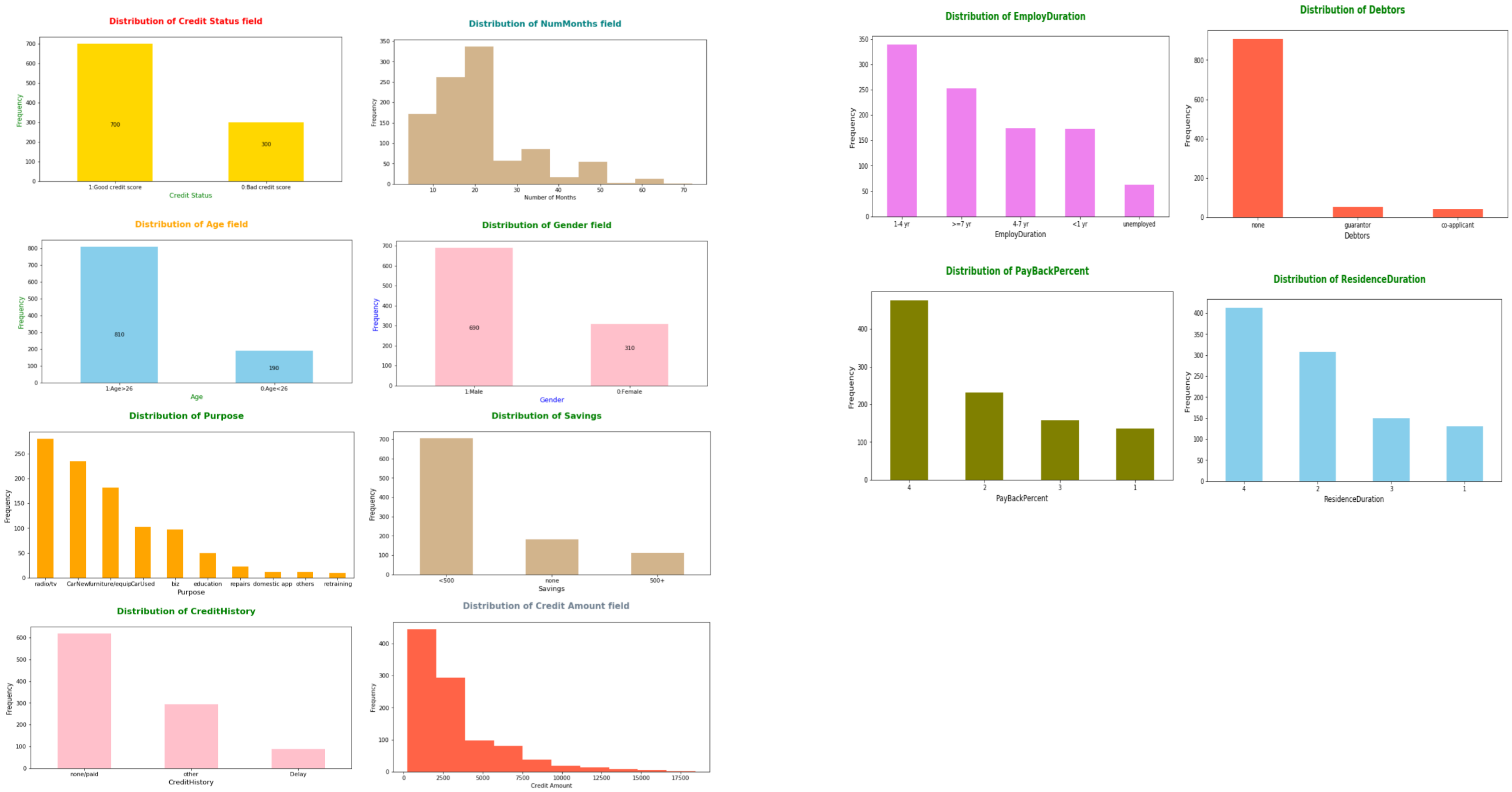


Fig) Visualizing target field (CreditStatus) and input fields

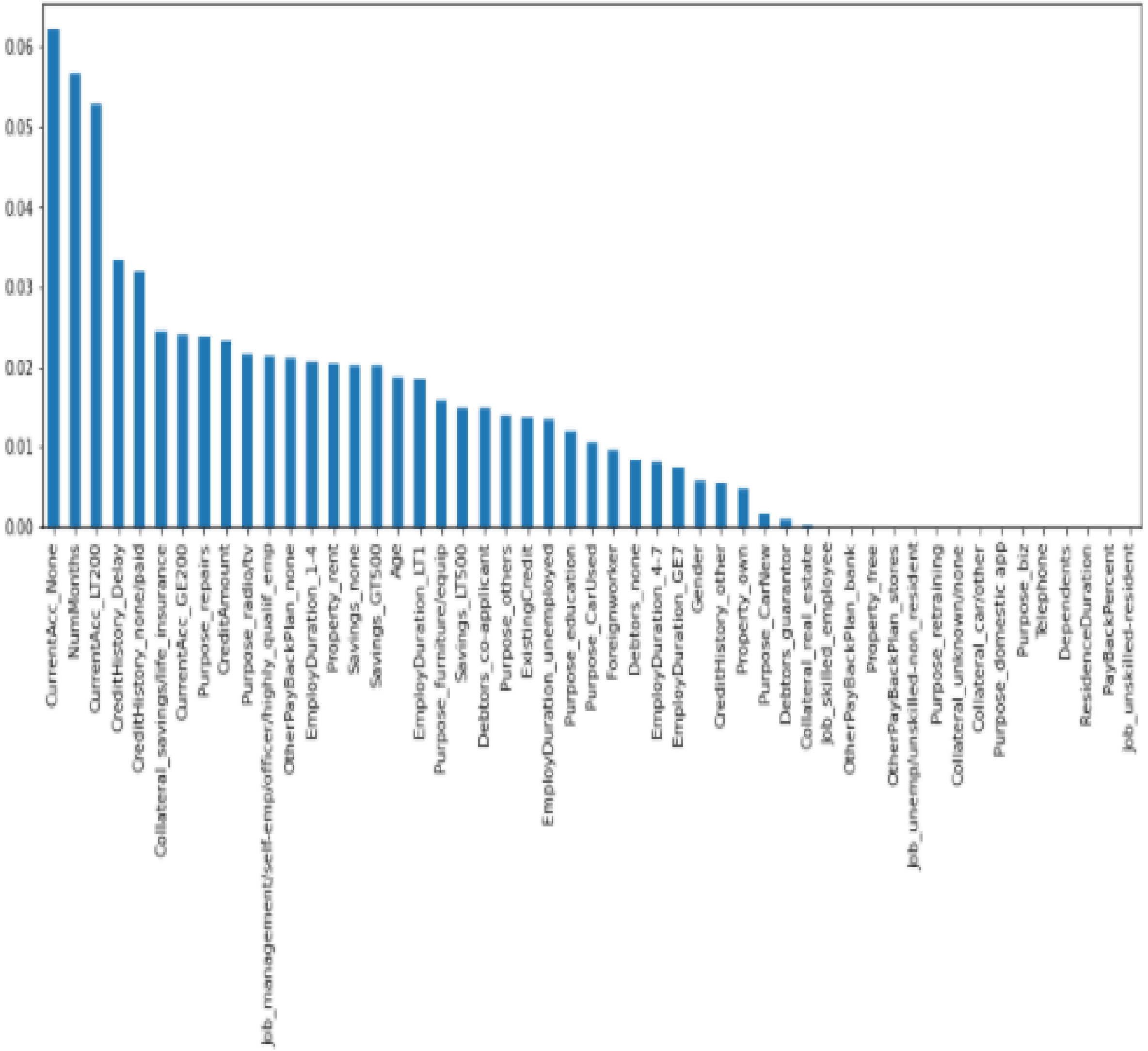


Fig) list of columns after dummy coding arranged in descending order of Mutual information of input field with target field

```
CurrentAcc_None                       uint8
NumMonths                             int64
CurrentAcc_LT200                      uint8
CreditHistory_Delay                   uint8
CreditHistory_none/paid               uint8
Collateral_savings/life_insurance     uint8
CurrentAcc_GE200                      uint8
Purpose_repairs                       uint8
CreditAmount                        float64
Purpose_radio/tv                      uint8
Gender                                int64
Age                                   int64
```

Fig) list of selected features for ML modeling

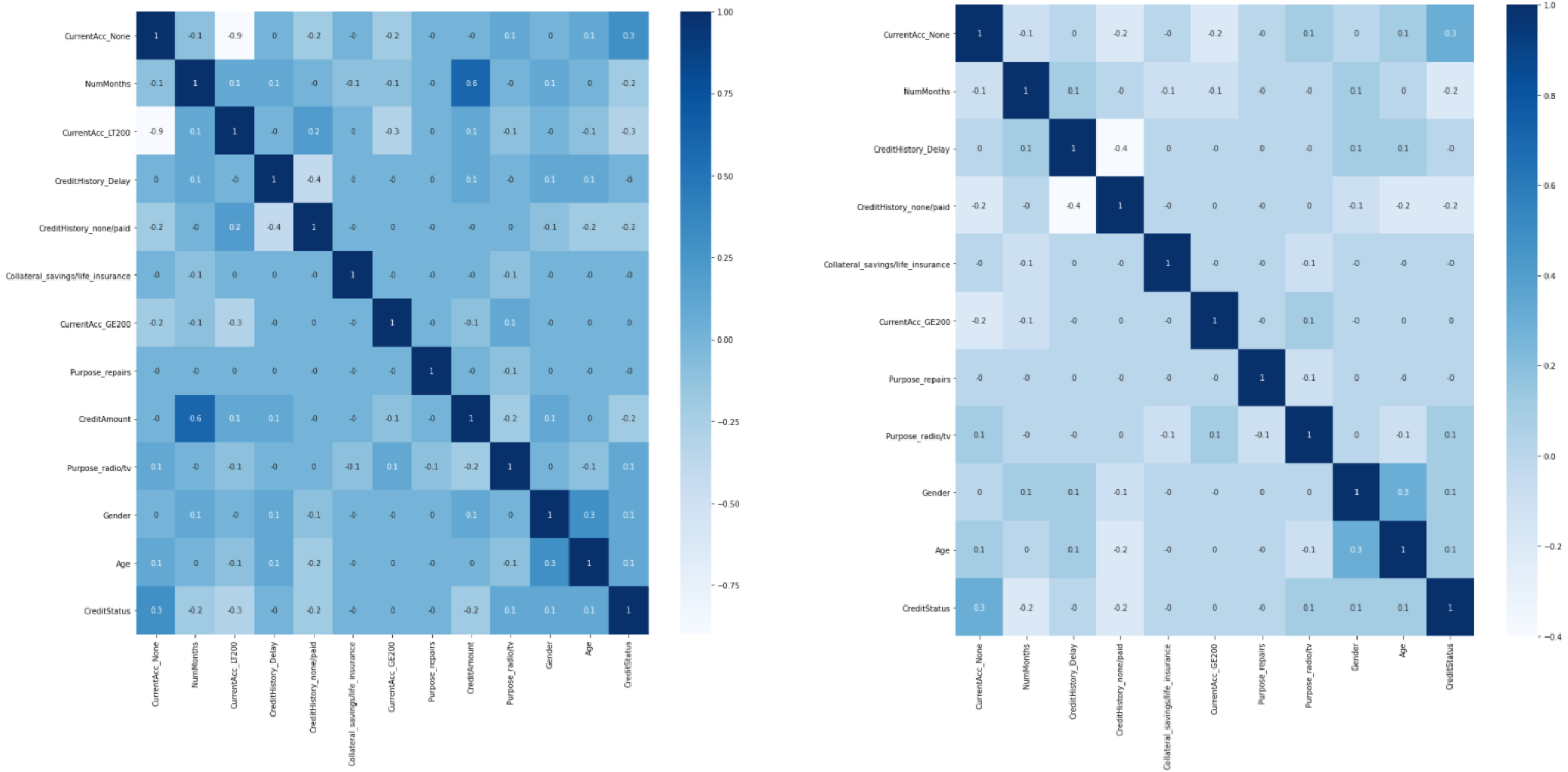


Fig) Correlation plot before and after removing multicollinearity (one (CreditAmount, CurrentAcc_LT200) of the highly correlated field pairs (CreditAmount, NumMonths ; CurrentAcc_None, CurrentAcc_LT200) is dropped)

Detect bias (mean difference)

Group mean difference method mean_difference() is used to identify privileged class. A negative value indicates less favourable outcomes for the unprivileged groups

Age:

```
Difference in mean outcomes between unprivileged and privileged groups = -0.179721
```

Age >26:1 (privileged class) is getting ~18% more positive outcome than unprevileged class.

Gender:

```
Difference in mean outcomes between unprivileged and privileged groups = -0.115809
```

Male (privileged class) is getting ~12% more positive outcome than unprivileged class.

Bias Mitigation Techniques:

- Pre-Processing Algorithms: do not change the model, only works on dataset before training
    - Reweighing: different weights are assigned to reduce effect of favouritism of a specified group.
    - Disparate Impact Remover (DIR): based on the concept of DI. It modifies the value of protected attribute to remove distinguishing factors
- In-Processing Algorithms: modify ml model
    - Adversarial Debiasing: introduces backward feedback(negative gradient) for predicting protected attribute which is achieved by using adversarial model that learns from difference between protected and other attributes.
    - Prejudice Remover Regularizer: if a model's decision is dependent on a protected attribute, it is called a direct prejudice. To handle this , we can remove this protected variable or regulate its effect on ml model. This regularization is used under this approach where a regularizer is implemented that computes the effect of protected attribute.
- Post-Processing Algorithms: modifies the predicted results instead of ml models or input data
    - Equalized odds (E): it changes the output labels to optimize EOD metric. A linear program is solved to obtain probabilities of modifying prediction.
    - Calibrated Equalized odds: this optimizes EOD metric by using calibrated prediction score produced by classifier.
    - Reject Option Classification: it favors the instances in privileged group over unprivileged ones that lie in the decision boundary with high uncertainty.

Measures of fairness used:

- Metrics based on base rates:
  - Disparate Impact (DI): ratio between the probability of unprivileged group gets favourable prediction and the probability of privileged group gets favourable prediction
  - Statistical Parity Difference (SPD): similar to DI but instead of ratios, differences is calculated
- Metrics based on group conditioned rates:
  - Equal Opportunity Difference (EOD): difference between TPR values of unprivileged and privileged groups.
  - Average Odds Difference (AOD): average of false positive rate difference between FPR of unprivileged and privileged groups and TPR of unprivileged and privileged groups.
  - Error Rate Difference (ERD):
    - Error rate ERR=FPR+FNR
    - ERD= ERR(U)-ERR(P)
- Metrics based on individual fairness:
  - Consistency (CNT): measures how similar are the predictions when the instances are similar.
  - Theil Index (TI) / Entropy Index: Measures both group and individual fairness.