Flight Predictions

Objective

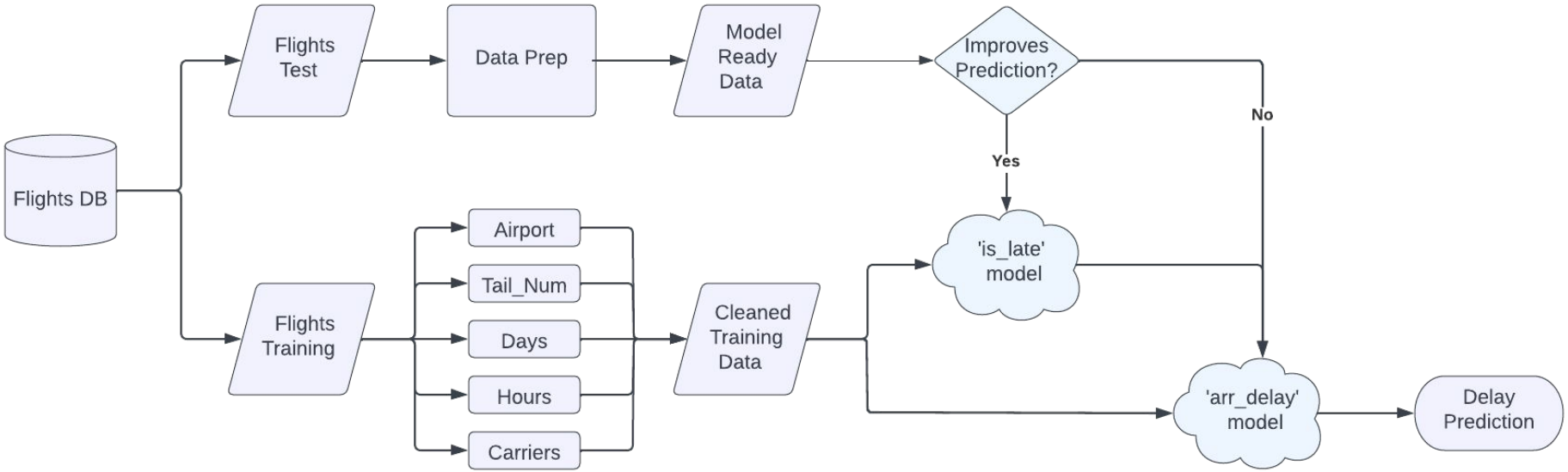
What:

Build a model to predict the arrival delays of US flights

Why:

- Operational Efficiency
- On Time Performance (OTP)
- Customer satisfaction

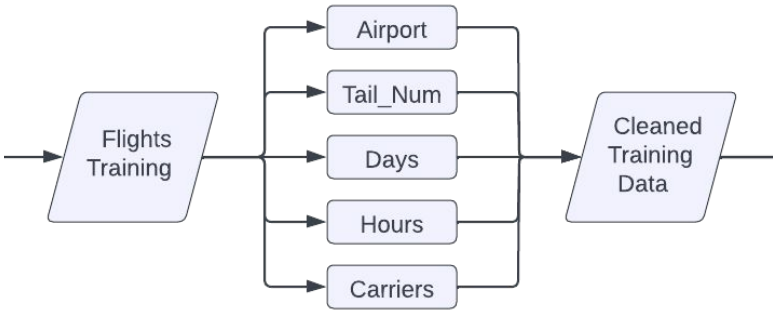
Data Pipeline

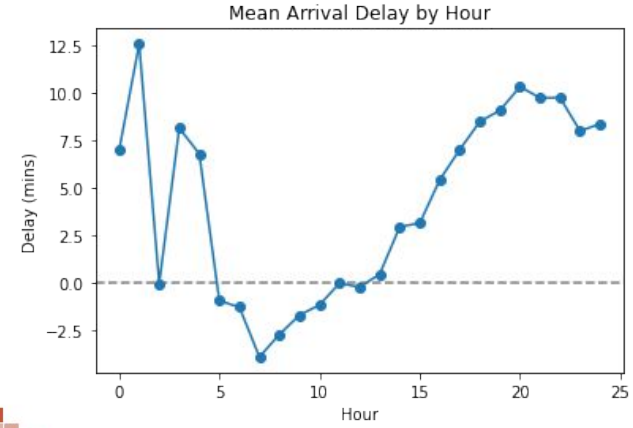


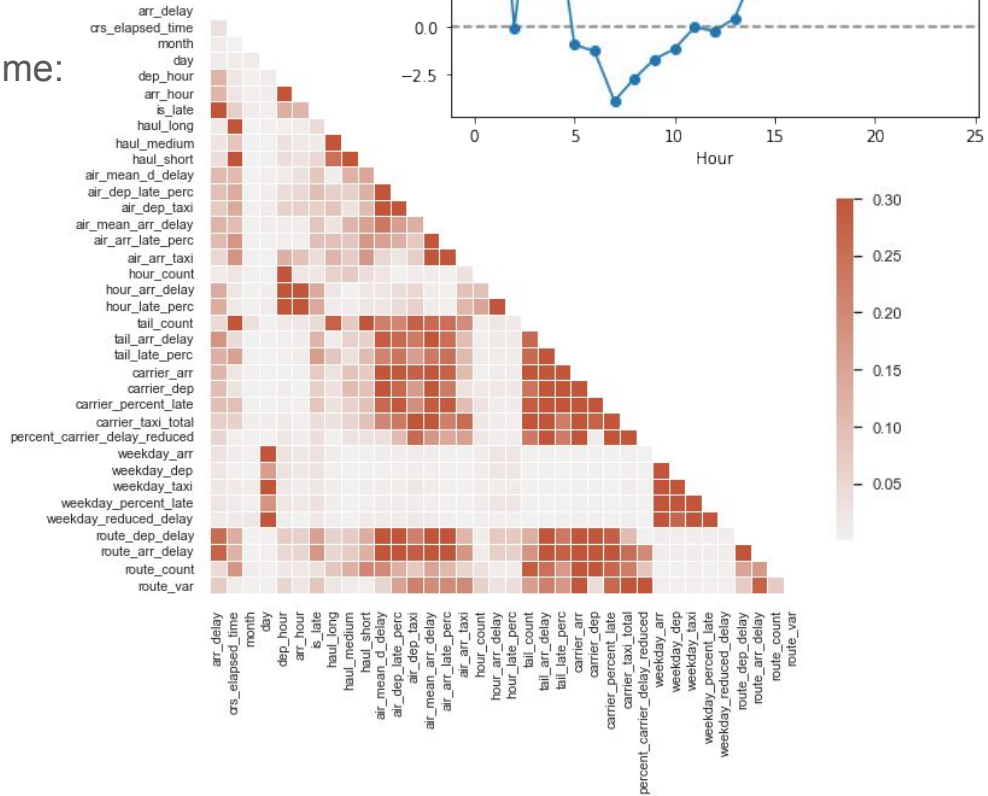
Feature Engineering

New features by airport, carrier, tail number, and time:

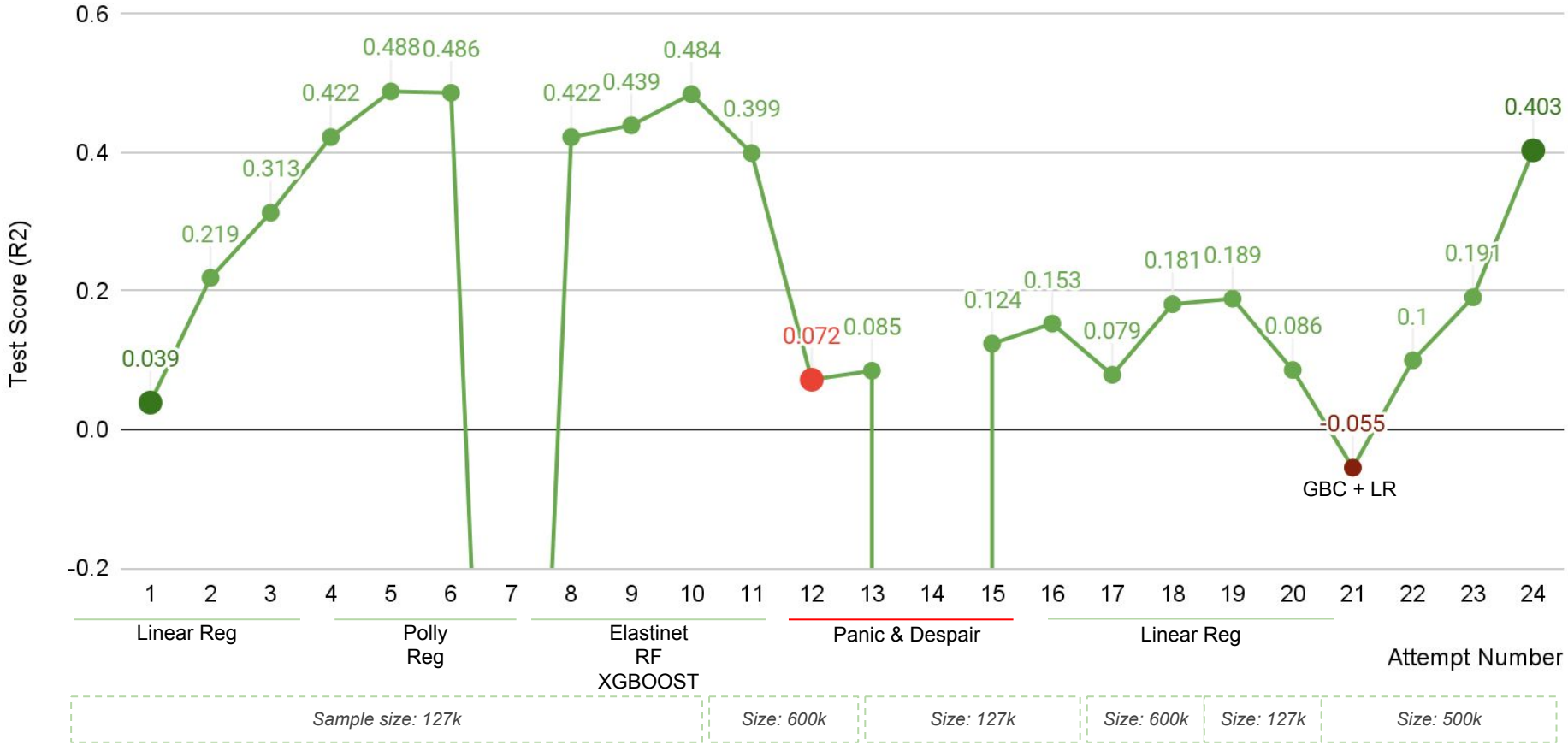
- Arrival delay %
- Arrival delay average
- Other: STD / Var / Counts







Test Score (R2) for Each Model Attempt

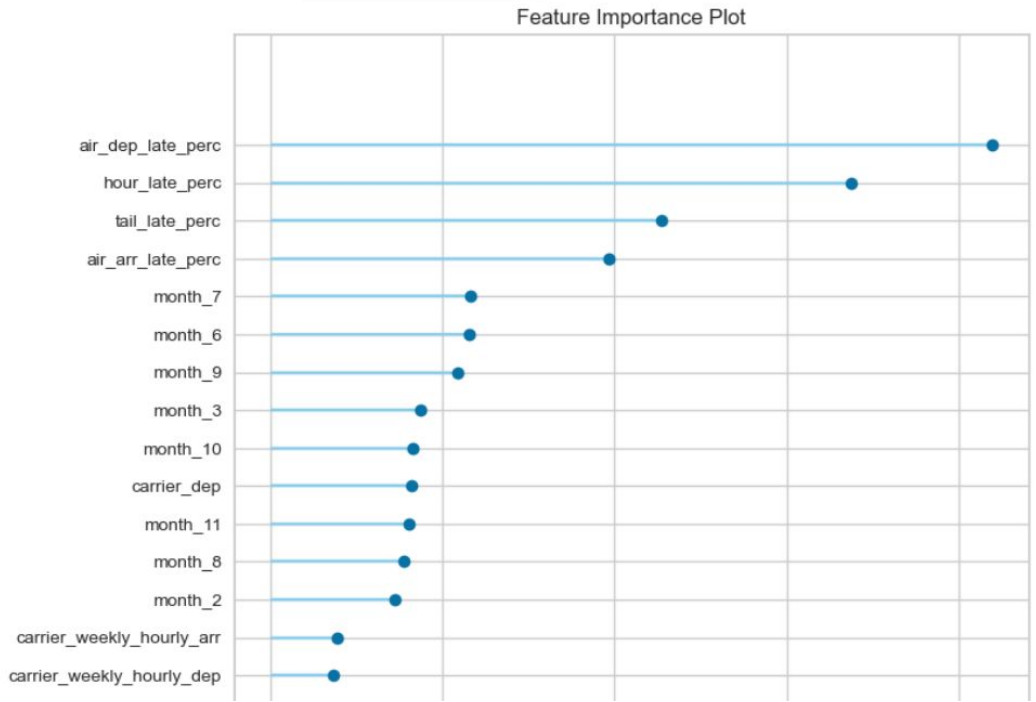


Business Impacts

- Base for more complex prediction model

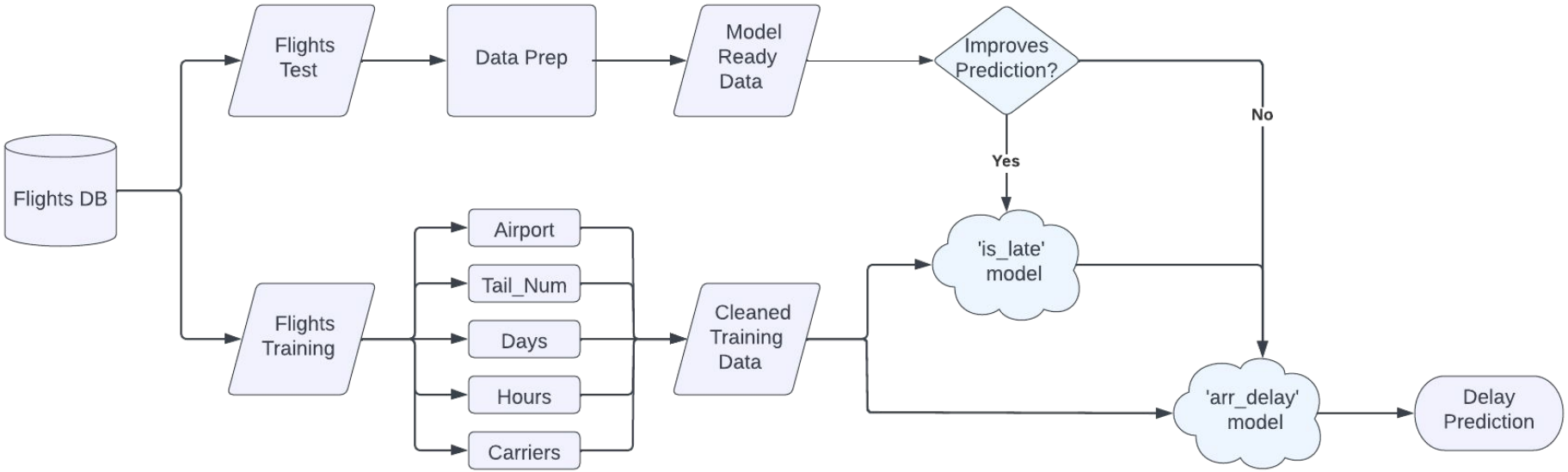
- Operational output
  - Scheduling to physical crew

- Communicative
  - Notify Airports
  - Customers
  - Staff

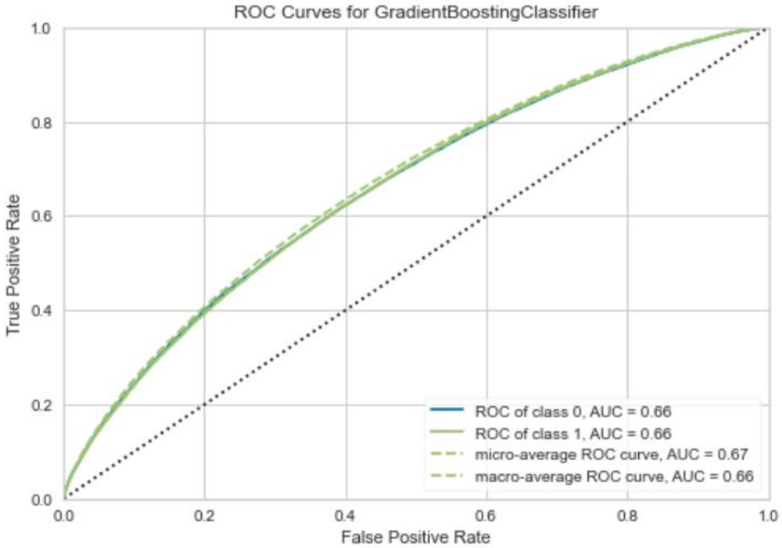


Conclusion

- Various Challenges: resource consideration
- Data & Feature Engineering **>** Models & Tuning
- Why? Improved Business Efficiency


Flights DB
Flights Test
Data Prep
Model Ready Data
Improves Prediction?
No
Yes
Flights Training
Airport
Tail_Num
Days
Hours
Carriers
Cleaned Training Data
'is_late' model
'arr_delay' model
Delay Prediction

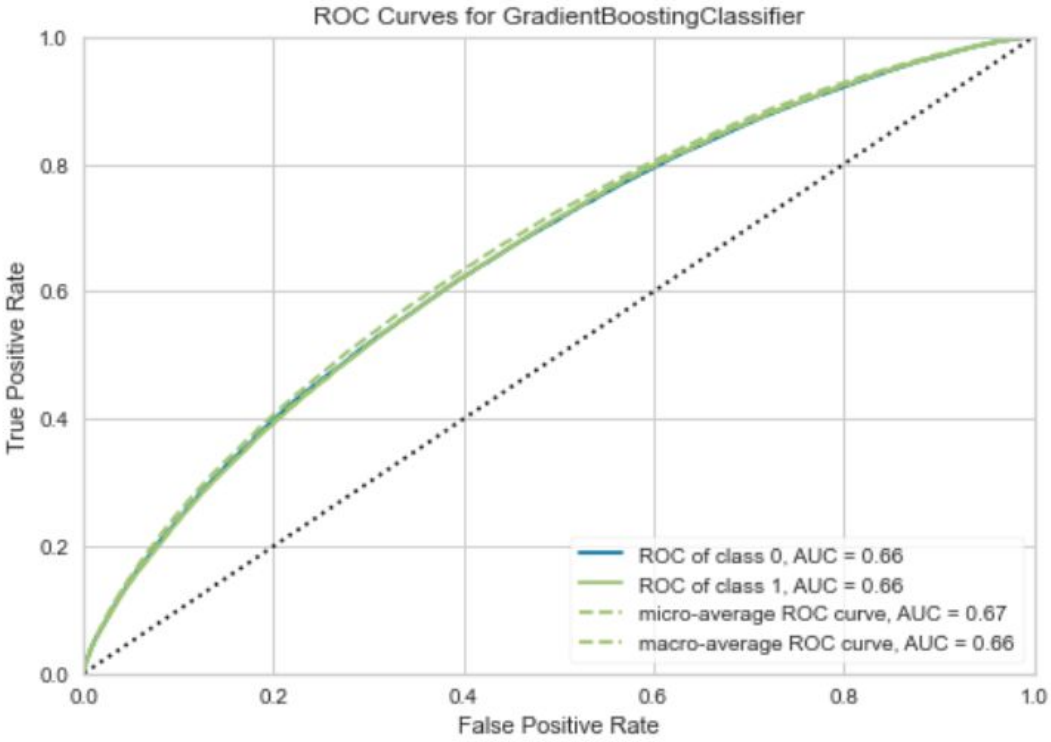
	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
ridge	Ridge Regression	18.3894	921.8518	30.3589	0.4080	1.1331	2.1645	0.3390
lr	Linear Regression	18.3944	922.0862	30.3628	0.4078	1.1335	2.1647	1.4420
br	Bayesian Ridge	18.3872	922.5576	30.3705	0.4075	1.1334	2.1635	2.1240
en	Elastic Net	18.3417	923.7750	30.3906	0.4067	1.1323	2.1469	1.3330
lasso	Lasso Regression	18.3339	924.1628	30.3970	0.4065	1.1321	2.1431	0.6950
omp	Orthogonal Matching Pursuit	18.4463	932.2639	30.5299	0.4013	1.1346	2.1641	0.2860
huber	Huber Regressor	17.7972	1071.1186	32.7174	0.3122	1.1723	1.5649	7.5610
par	Passive Aggressive Regressor	23.1458	1118.9860	33.3849	0.2822	1.2298	3.2862	0.4140
knn	K Neighbors Regressor	19.9338	1170.7663	34.2126	0.2482	1.2571	2.1578	7.7530
llar	Lasso Least Angle Regression	24.0670	1557.5349	39.4610	-0.0001	1.0747	2.4806	0.7670
dt	Decision Tree Regressor	24.9717	2011.4735	44.8472	-0.2922	1.3187	3.0135	2.7390



Split	Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
CV-Train	0	0.6199	0.6616	0.4991	0.5974	0.5438	0.2225	0.2252
	1	0.6189	0.6610	0.4956	0.5967	0.5415	0.2202	0.2230
	2	0.6199	0.6616	0.4974	0.5978	0.5430	0.2223	0.2250
	3	0.6200	0.6611	0.4964	0.5983	0.5426	0.2224	0.2253
	4	0.6197	0.6607	0.4918	0.5988	0.5401	0.2211	0.2242
	5	0.6201	0.6619	0.4963	0.5984	0.5426	0.2226	0.2255
	6	0.6202	0.6615	0.4946	0.5989	0.5418	0.2224	0.2254
	7	0.6201	0.6616	0.4973	0.5981	0.5430	0.2226	0.2254
	8	0.6196	0.6616	0.4984	0.5971	0.5433	0.2218	0.2245
	9	0.6201	0.6615	0.4969	0.5983	0.5429	0.2227	0.2255
CV-Val	0	0.6150	0.6548	0.4942	0.5909	0.5383	0.2127	0.2152
	1	0.6180	0.6594	0.4927	0.5958	0.5394	0.2180	0.2209
	2	0.6120	0.6543	0.4889	0.5873	0.5336	0.2061	0.2086
	3	0.6217	0.6626	0.5032	0.5993	0.5471	0.2265	0.2291
	4	0.6235	0.6642	0.4945	0.6043	0.5439	0.2288	0.2321
	5	0.6178	0.6529	0.4915	0.5959	0.5387	0.2176	0.2205
	6	0.6174	0.6546	0.4932	0.5948	0.5393	0.2170	0.2198
	7	0.6132	0.6521	0.4904	0.5889	0.5351	0.2087	0.2113
	8	0.6136	0.6541	0.4907	0.5894	0.5356	0.2094	0.2120
	9	0.6154	0.6566	0.4915	0.5920	0.5371	0.2129	0.2156
CV-Train	Mean	0.6199	0.6614	0.4964	0.5980	0.5425	0.2221	0.2249
	Std	0.0004	0.0003	0.0020	0.0007	0.0010	0.0008	0.0008
CV-Val	Mean	0.6168	0.6566	0.4931	0.5939	0.5388	0.2158	0.2185
	Std	0.0035	0.0039	0.0038	0.0050	0.0039	0.0071	0.0072
Train	nan	0.6193	0.6609	0.4955	0.5973	0.5417	0.2209	0.2237

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
gbc	Gradient Boosting Classifier	0.6181	0.6580	0.4891	0.5953	0.5369	0.2173	0.2204	125.1610
lda	Linear Discriminant Analysis	0.6162	0.6536	0.4625	0.5985	0.5218	0.2099	0.2148	2.8500
ridge	Ridge Classifier	0.6161	0.0000	0.4614	0.5987	0.5212	0.2097	0.2147	0.3680
ada	Ada Boost Classifier	0.6130	0.6498	0.4771	0.5898	0.5275	0.2060	0.2093	27.7150
lr	Logistic Regression	0.6026	0.6356	0.4158	0.5861	0.4865	0.1774	0.1843	11.1280
rf	Random Forest Classifier	0.5968	0.6284	0.4775	0.5646	0.5174	0.1753	0.1771	47.9910
et	Extra Trees Classifier	0.5954	0.6208	0.4756	0.5628	0.5155	0.1723	0.1741	62.6270
nb	Naive Bayes	0.5886	0.6127	0.4571	0.5554	0.5015	0.1568	0.1590	0.3210
knn	K Neighbors Classifier	0.5599	0.5712	0.4727	0.5152	0.4930	0.1054	0.1057	12.2070
dt	Decision Tree Classifier	0.5396	0.5359	0.4963	0.4916	0.4939	0.0716	0.0716	7.7380
svm	SVM - Linear Kernel	0.5349	0.0000	0.5007	0.5752	0.3890	0.0642	0.1008	67.9480
qda	Quadratic Discriminant Analysis	0.4971	0.4987	0.5151	0.4503	0.4449	-0.0029	-0.0024	1.7180

Binary Classification Model



Split	Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
CV-Train	0	0.6199	0.6616	0.4991	0.5974	0.5438	0.2225	0.2252
	1	0.6189	0.6610	0.4956	0.5967	0.5415	0.2202	0.2230
	2	0.6199	0.6616	0.4974	0.5978	0.5430	0.2223	0.2250
	3	0.6200	0.6611	0.4964	0.5983	0.5426	0.2224	0.2253
	4	0.6197	0.6607	0.4918	0.5988	0.5401	0.2211	0.2242
	5	0.6201	0.6619	0.4963	0.5984	0.5426	0.2226	0.2255
	6	0.6202	0.6615	0.4946	0.5989	0.5418	0.2224	0.2254
	7	0.6201	0.6616	0.4973	0.5981	0.5430	0.2226	0.2254
	8	0.6196	0.6616	0.4984	0.5971	0.5433	0.2218	0.2245
	9	0.6201	0.6615	0.4969	0.5983	0.5429	0.2227	0.2255
CV-Val	0	0.6150	0.6548	0.4942	0.5909	0.5383	0.2127	0.2152
	1	0.6180	0.6594	0.4927	0.5958	0.5394	0.2180	0.2209
	2	0.6120	0.6543	0.4889	0.5873	0.5336	0.2061	0.2086
	3	0.6217	0.6626	0.5032	0.5993	0.5471	0.2265	0.2291
	4	0.6235	0.6642	0.4945	0.6043	0.5439	0.2288	0.2321
	5	0.6178	0.6529	0.4915	0.5959	0.5387	0.2176	0.2205
	6	0.6174	0.6546	0.4932	0.5948	0.5393	0.2170	0.2198
	7	0.6132	0.6521	0.4904	0.5889	0.5351	0.2087	0.2113
	8	0.6136	0.6541	0.4907	0.5894	0.5356	0.2094	0.2120
	9	0.6154	0.6566	0.4915	0.5920	0.5371	0.2129	0.2156
CV-Train	Mean	0.6199	0.6614	0.4964	0.5980	0.5425	0.2221	0.2249
	Std	0.0004	0.0003	0.0020	0.0007	0.0010	0.0008	0.0008
CV-Val	Mean	0.6168	0.6568	0.4931	0.5932	0.5388	0.2168	0.2198
	Std	0.0035	0.0039	0.0038	0.0050	0.0039	0.0071	0.0072
Train	nan	0.6193	0.6609	0.4955	0.5973	0.5417	0.2209	0.2237