

## TELCO CUSTOMER CHURN: IBM SAMPLE DATASET

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	\
0	7590-VHVEG	Female	0	Yes	No	1	No	
1	5575-GNVDE	Male	0	No	No	34	Yes	
2	3668-QPYBK	Male	0	No	No	2	Yes	
3	7795-CFOCW	Male	0	No	No	45	No	
4	9237-HQITU	Female	0	No	No	2	Yes	

	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	\
0	No phone service	DSL	No	...	No	
1	No	DSL	Yes	...	Yes	
2	No	DSL	Yes	...	No	
3	No phone service	DSL	Yes	...	Yes	
4	No	Fiber optic	No	...	No	

	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	\
0	No	No	No	Month-to-month	Yes	
1	No	No	No	One year	No	
2	No	No	No	Month-to-month	Yes	
3	Yes	No	No	One year	No	
4	No	No	No	Month-to-month	Yes	

	PaymentMethod	MonthlyCharges	TotalCharges	Churn
0	Electronic check	29.85	29.85	No
1	Mailed check	56.95	1889.5	No
2	Mailed check	53.85	108.15	Yes
3	Bank transfer (automatic)	42.30	1840.75	No
4	Electronic check	70.70	151.65	Yes

[5 rows x 21 columns]

-----

CHECK FOR COLUMN DATATYPES:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 7043 entries, 0 to 7042

Data columns (total 21 columns):

#	Column	Non-Null	Count	Dtype
0	customerID	7043	non-null	object
1	gender	7043	non-null	object
2	SeniorCitizen	7043	non-null	int64
3	Partner	7043	non-null	object
4	Dependents	7043	non-null	object
5	tenure	7043	non-null	int64
6	PhoneService	7043	non-null	object
7	MultipleLines	7043	non-null	object
8	InternetService	7043	non-null	object
9	OnlineSecurity	7043	non-null	object
10	OnlineBackup	7043	non-null	object
11	DeviceProtection	7043	non-null	object
12	TechSupport	7043	non-null	object
13	StreamingTV	7043	non-null	object
14	StreamingMovies	7043	non-null	object
15	Contract	7043	non-null	object
16	PaperlessBilling	7043	non-null	object
17	PaymentMethod	7043	non-null	object
18	MonthlyCharges	7043	non-null	float64
19	TotalCharges	7043	non-null	object
20	Churn	7043	non-null	object

dtypes: float64(1), int64(2), object(18)

memory usage: 1.1+ MB

'TotalCharges' data type is string, which is weird.

Blank strings might be the reason why it cannot be converted to float.

-----

CHECK FOR DATAFRAME TOTAL NULL VALUES:

customerID	0
gender	0
SeniorCitizen	0
Partner	0
Dependents	0
tenure	0
PhoneService	0
MultipleLines	0
InternetService	0
OnlineSecurity	0
OnlineBackup	0
DeviceProtection	0
TechSupport	0
StreamingTV	0
StreamingMovies	0
Contract	0
PaperlessBilling	0
PaymentMethod	0
MonthlyCharges	0
TotalCharges	0
Churn	0

dtype: int64

NOTE: Blank strings can be classified as not null.

-----

Blank strings in 'TotalCharges' are replaced with NaN

CHECK FOR DATAFRAME TOTAL NULL VALUES:

customerID	0
gender	0
SeniorCitizen	0
Partner	0
Dependents	0
tenure	0
PhoneService	0
MultipleLines	0
InternetService	0
OnlineSecurity	0
OnlineBackup	0
DeviceProtection	0
TechSupport	0
StreamingTV	0
StreamingMovies	0
Contract	0
PaperlessBilling	0
PaymentMethod	0
MonthlyCharges	0
TotalCharges	11
Churn	0
dtype: int64	
Missing TotalCharges: 11 (less missing data are safe to drop)	
-----	

'TotalCharges' datatype is converted to float

CHECK FOR COLUMN DATATYPES:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 7043 entries, 0 to 7042

Data columns (total 21 columns):

#	Column	Non-Null Count	Dtype
0	customerID	7043 non-null	object
1	gender	7043 non-null	object
2	SeniorCitizen	7043 non-null	int64
3	Partner	7043 non-null	object
4	Dependents	7043 non-null	object
5	tenure	7043 non-null	int64
6	PhoneService	7043 non-null	object
7	MultipleLines	7043 non-null	object
8	InternetService	7043 non-null	object
9	OnlineSecurity	7043 non-null	object
10	OnlineBackup	7043 non-null	object
11	DeviceProtection	7043 non-null	object
12	TechSupport	7043 non-null	object
13	StreamingTV	7043 non-null	object
14	StreamingMovies	7043 non-null	object
15	Contract	7043 non-null	object
16	PaperlessBilling	7043 non-null	object
17	PaymentMethod	7043 non-null	object
18	MonthlyCharges	7043 non-null	float64
19	TotalCharges	7032 non-null	float64
20	Churn	7043 non-null	object

dtypes: float64(2), int64(2), object(17)

memory usage: 1.1+ MB

-----

CHECK FOR DATAFRAME TOTAL NULL VALUES:

customerID	0
gender	0
SeniorCitizen	0
Partner	0
Dependents	0
tenure	0
PhoneService	0
MultipleLines	0
InternetService	0
OnlineSecurity	0
OnlineBackup	0
DeviceProtection	0
TechSupport	0
StreamingTV	0
StreamingMovies	0
Contract	0
PaperlessBilling	0
PaymentMethod	0
MonthlyCharges	0
TotalCharges	0
Churn	0

dtype: int64

No Null values found

-----

'customerID' dropped

-----

Redundancy reworked:

- 'MultipleLines': 'No phone service' is set to 'No'
  - 'OnlineSecurity': 'No internet service' is set to 'No'
  - 'OnlineBackup': 'No internet service' is set to 'No'
  - 'DeviceProtection': 'No internet service' is set to 'No'
  - 'TechSupport': 'No internet service' is set to 'No'
  - 'StreamingTV': 'No internet service' is set to 'No'
  - 'StreamingMovies': 'No internet service' is set to 'No'
  - 'PaymentMethod': 'Bank transfer (automatic)' is set to 'Bank Transfer'
  - 'Credit card (automatic)' is set to 'Credit Card'
  - 'Contract': 'Month-to-month' is set to 'Month-To-Month',
  - 'Contract': 'One year' is set to 'One Year',
  - 'Contract': 'Two year' is set to 'Two Year'
-

Before SMOTE:

Churn

0 4130

1 1495

Name: count, dtype: int64

---

Afer SMOTE:

Churn

0 4130

1 4130

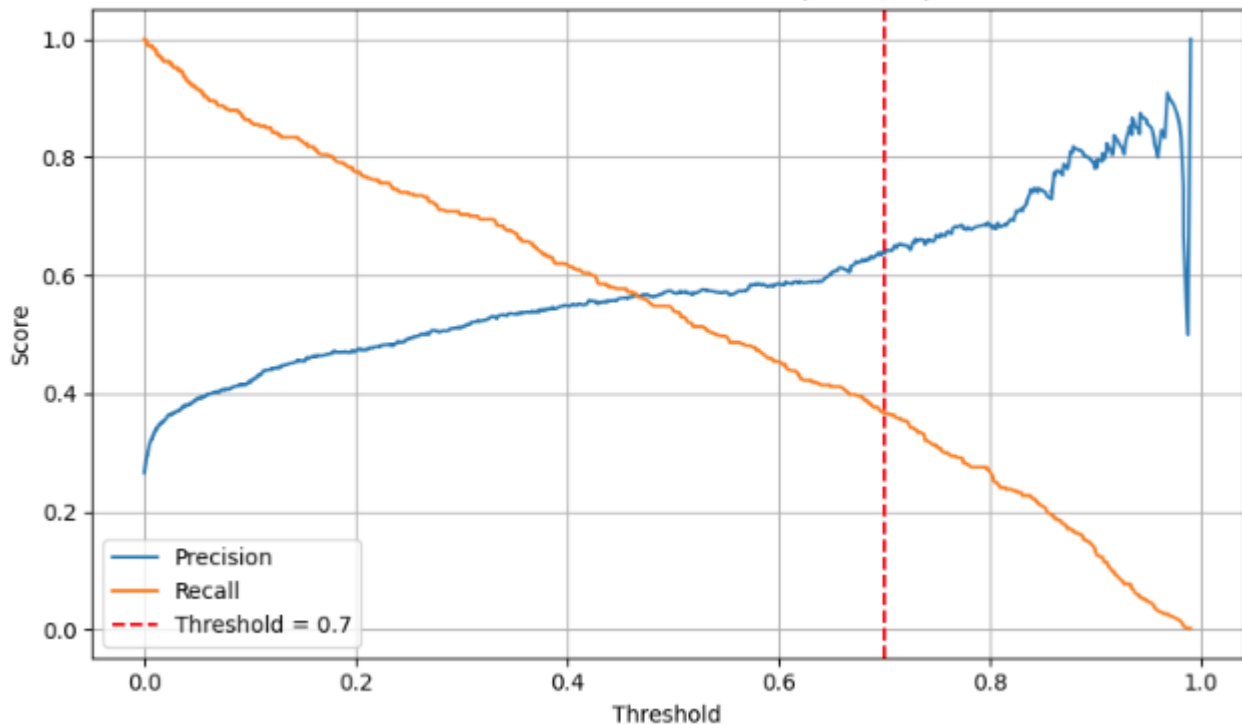
Name: count, dtype: int64

---

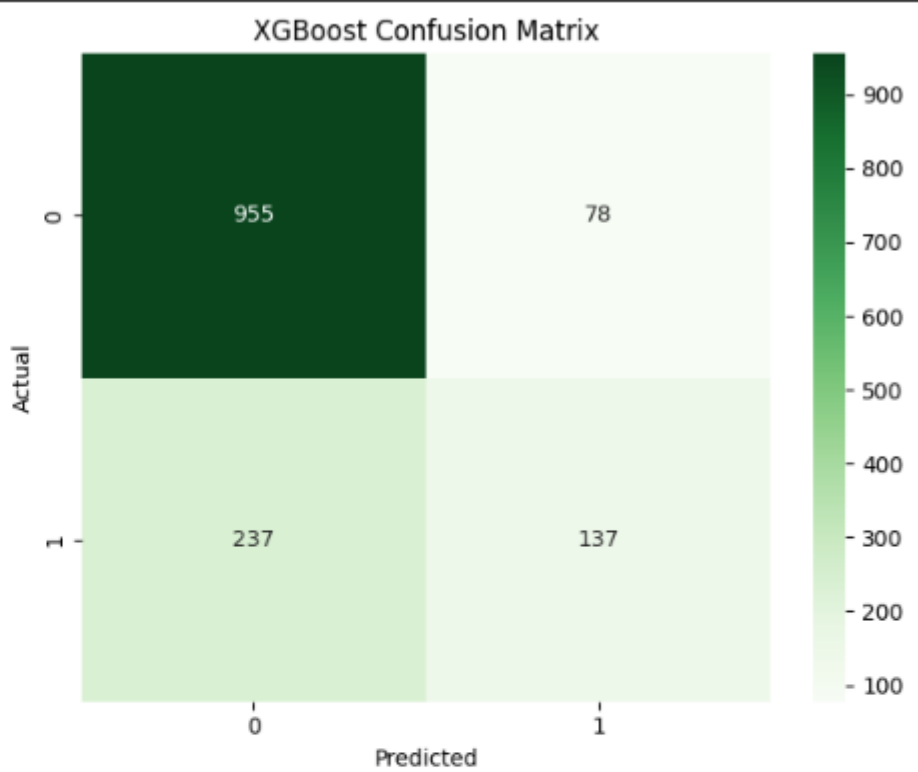
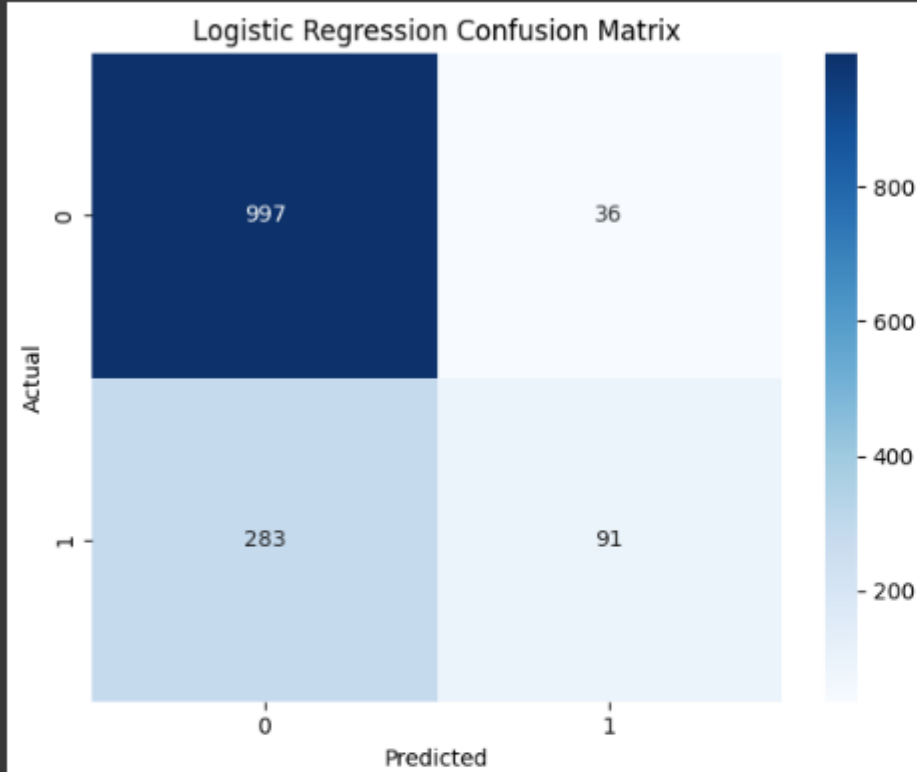


```
XGBClassifier(base_score=None, booster=None, callbacks=None,  
              colsample_bylevel=None, colsample_bynode=None,  
              colsample_bytrees=None, device=None, early_stopping_rounds=None,  
              enable_categorical=False, eval_metric='logloss',  
              feature_types=None, gamma=None, grow_policy=None,  
              importance_type=None, interaction_constraints=None,  
              learning_rate=None, max_bin=None, max_cat_threshold=None,  
              max_cat_to_onehot=None, max_delta_step=None, max_depth=None,  
              max_leaves=None, min_child_weight=None, missing=nan,  
              monotone_constraints=None, multi_strategy=None, n_estimators=None,  
              n_jobs=None, num_parallel_tree=None, random_state=None, ...)
```

Precision-Recall vs Threshold (XGBoost)



XGBoost Accuracy:  
0.7761194029850746



Logistic Regression Classification Report (threshold=0.7):

	precision	recall	f1-score	support
0	0.78	0.97	0.86	1033
1	0.72	0.24	0.36	374
accuracy			0.77	1407
macro avg	0.75	0.60	0.61	1407
weighted avg	0.76	0.77	0.73	1407

XGBoost Classification Report (threshold=0.7):

	precision	recall	f1-score	support
0	0.80	0.92	0.86	1033
1	0.64	0.37	0.47	374
accuracy			0.78	1407
macro avg	0.72	0.65	0.66	1407
weighted avg	0.76	0.78	0.75	1407

## TOP 10 PREDICTIONS

	Actual	LogReg_Predicted	LogReg_Probability	XGB_Predicted	\
0	0	0	0.014116	0	
1	0	0	0.616168	0	
2	0	0	0.005123	0	
3	1	0	0.296936	0	
4	0	0	0.109429	0	
5	1	0	0.513419	0	
6	0	0	0.024770	0	
7	0	0	0.240843	0	
8	1	0	0.681514	1	
9	0	0	0.012032	0	

## XGB\_Probability

0	0.005628
1	0.668626
2	0.021080
3	0.108084
4	0.546587
5	0.456047
6	0.049399
7	0.103120
8	0.715994
9	0.003061

-----

Predictions saved to churn\_predictions\_threshold\_0.7.csv

-----

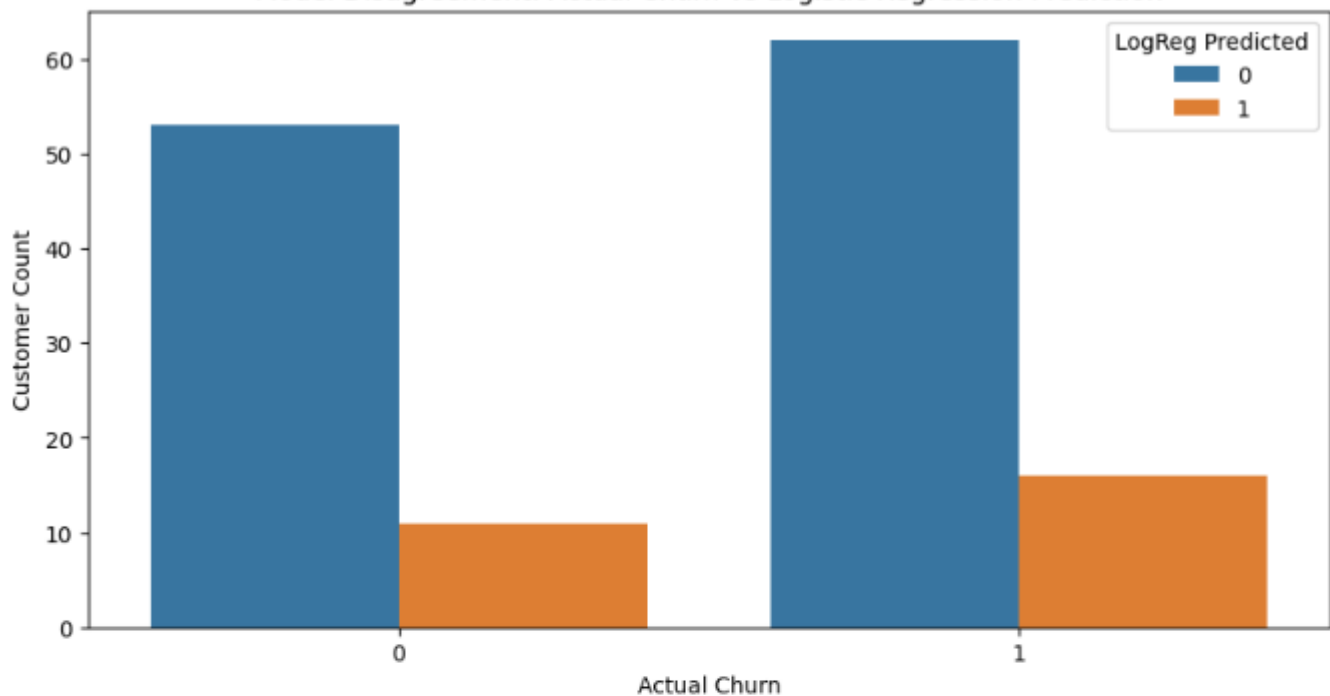
## Customers with model disagreement:

	Actual	LogReg_Predicted	LogReg_Probability	XGB_Predicted	\
8	1	0	0.681514	1	
10	0	1	0.741091	0	
12	0	0	0.690320	1	
32	0	0	0.386750	1	
53	1	0	0.624098	1	
76	0	0	0.491503	1	
91	1	1	0.710360	0	
107	0	0	0.621353	1	
108	0	1	0.716069	0	
109	1	0	0.597525	1	

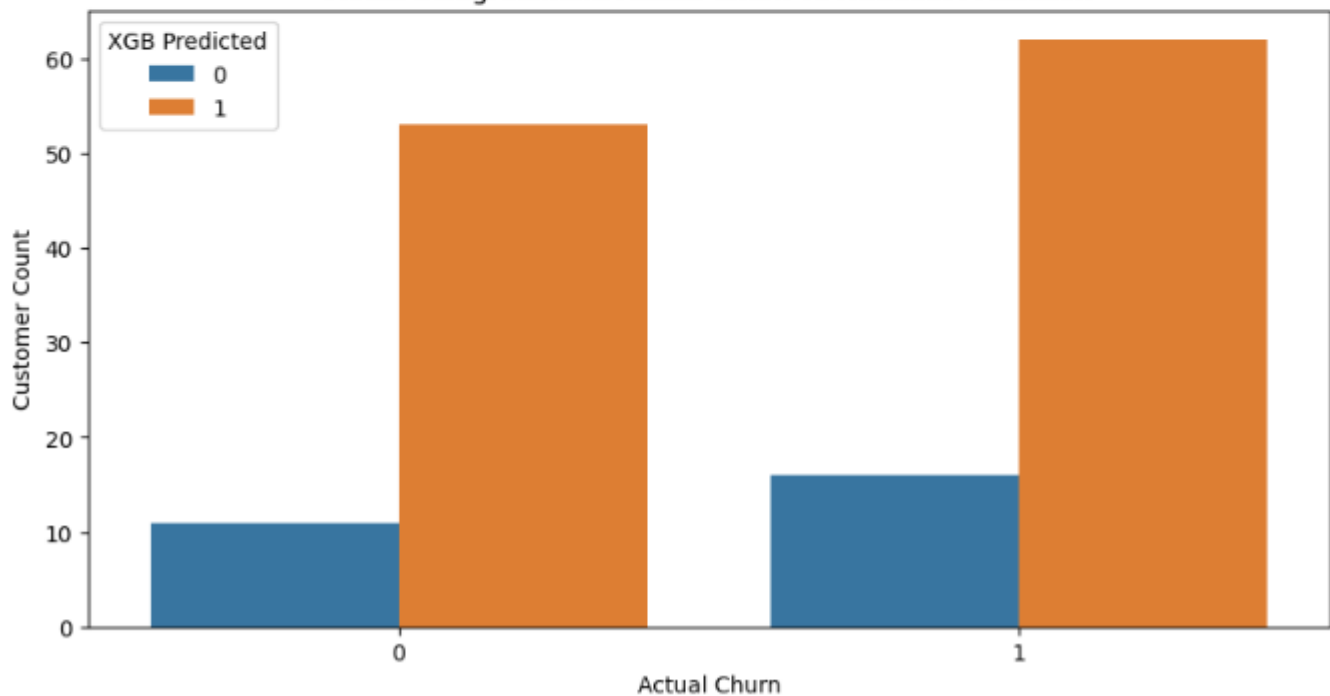
## XGB\_Probability

8	0.715994
10	0.638119
12	0.878543
32	0.908299
53	0.742643
76	0.810211
91	0.661225
107	0.934654
108	0.666903
109	0.825495

Model Disagreement: Actual Churn vs Logistic Regression Prediction



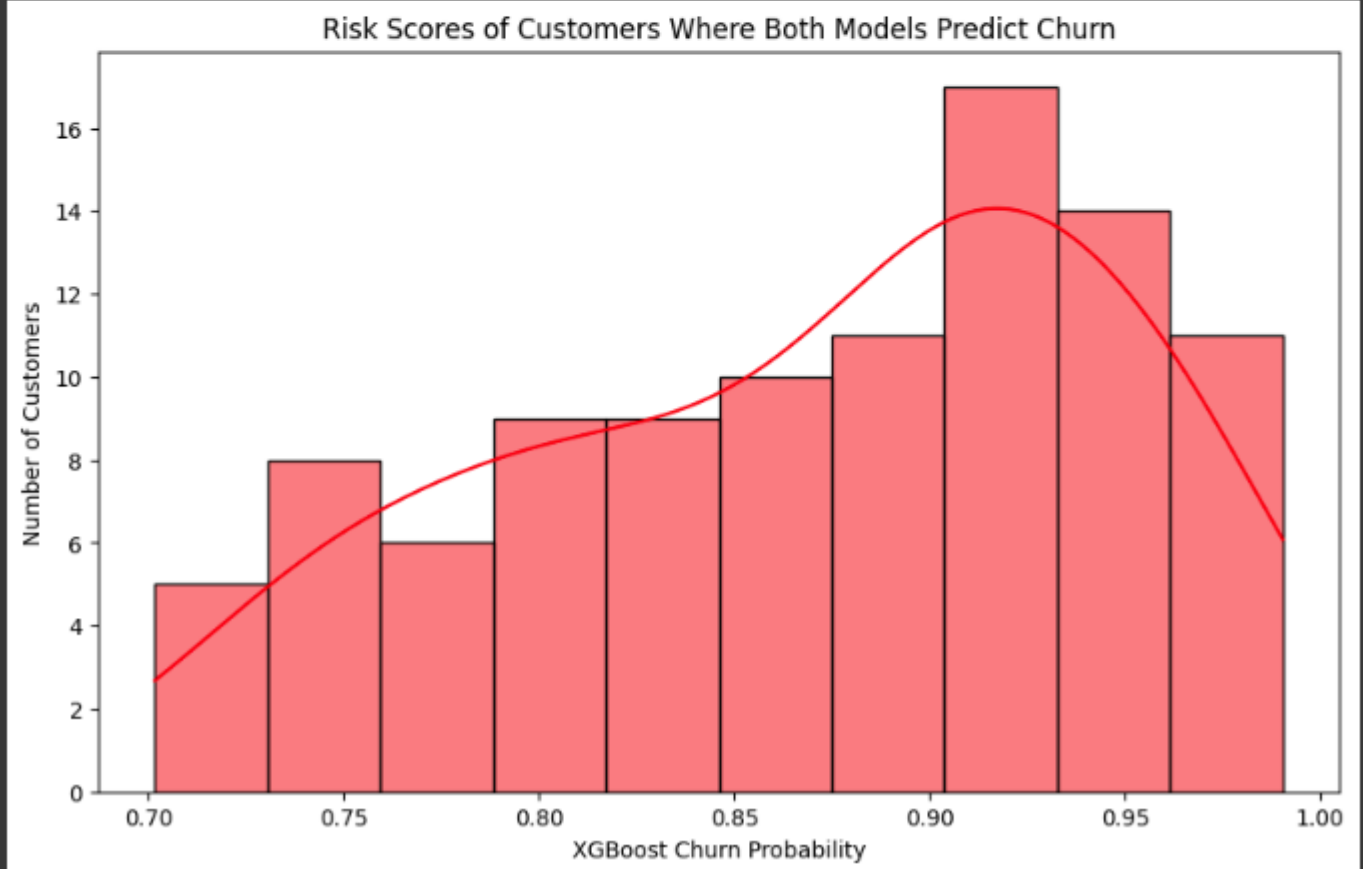
Model Disagreement: Actual Churn vs XGBoost Prediction



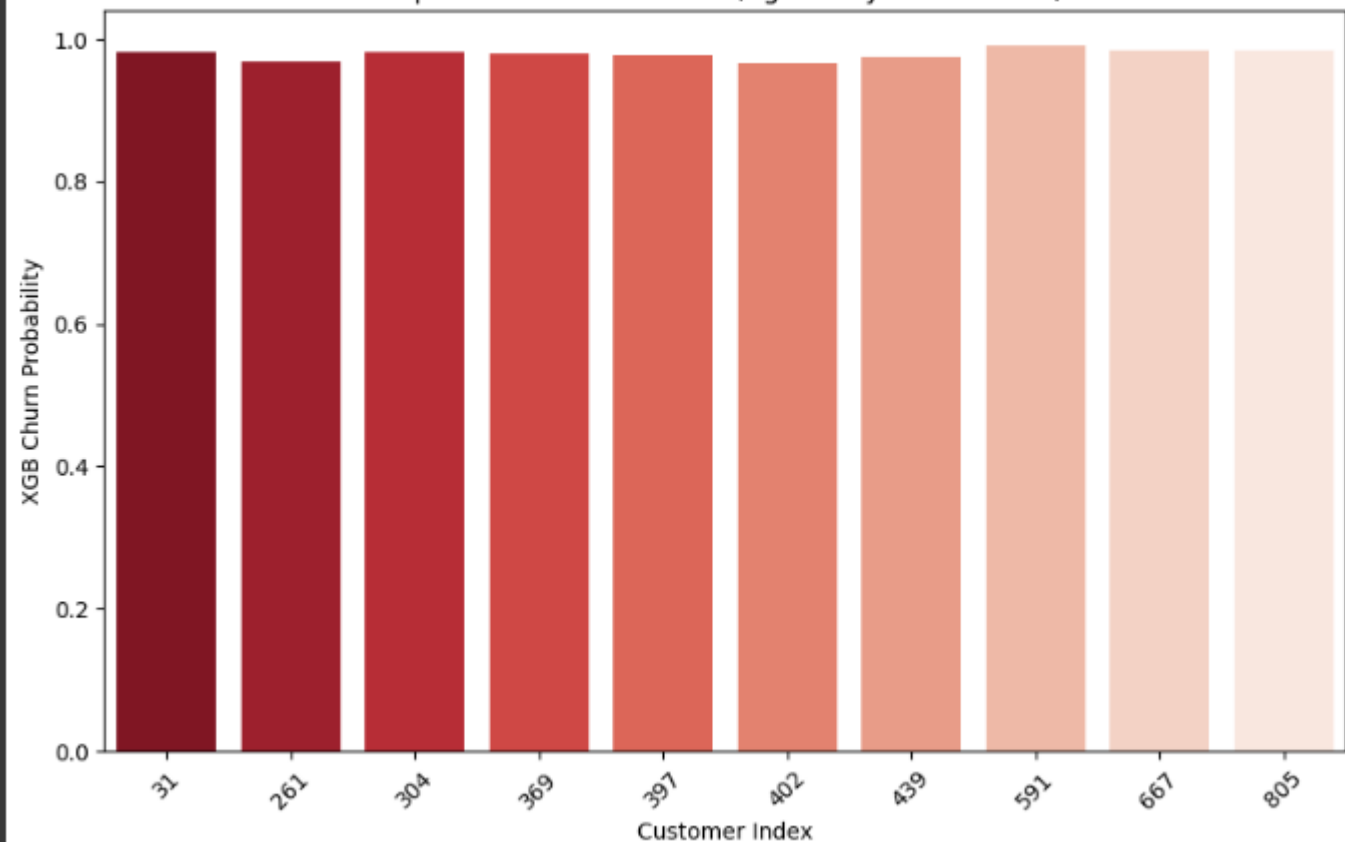
Customers with high-confidence churn predictions (both models agree):

	Actual	LogReg_Predicted	LogReg_Probability	XGB_Predicted	\
591	1	1	0.827168	1	
805	1	1	0.793616	1	
667	1	1	0.748259	1	
31	1	1	0.755957	1	
304	1	1	0.754535	1	
369	1	1	0.781424	1	
397	1	1	0.776491	1	
439	1	1	0.814610	1	
261	1	1	0.757458	1	
402	0	1	0.735962	1	

	XGB_Probability
591	0.990364
805	0.983712
667	0.983559
31	0.981304
304	0.981304
369	0.979757
397	0.977343
439	0.974331
261	0.968428
402	0.965864



Top 10 At-Risk Customers (Agreed by Both Models)



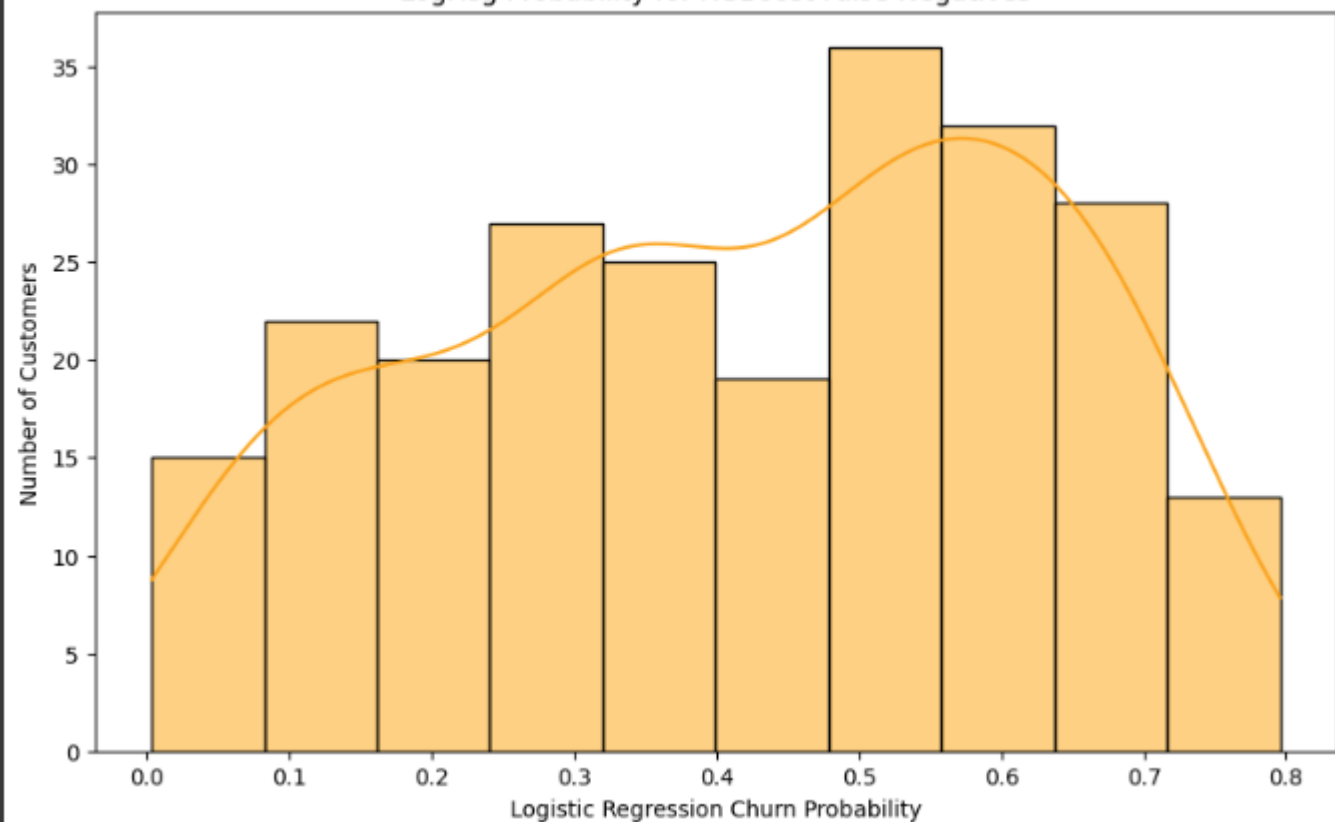
False negatives (actual churn but model missed):

	Actual	LogReg_Predicted	LogReg_Probability	XGB_Predicted	\
3	1	0	0.296936	0	
5	1	0	0.513419	0	
14	1	0	0.171549	0	
16	1	0	0.657338	0	
17	1	0	0.595503	0	
23	1	0	0.153976	0	
25	1	0	0.412559	0	
29	1	0	0.309959	0	
38	1	0	0.304124	0	
73	1	0	0.371406	0	

	XGB_Probability
3	0.108084
5	0.456047
14	0.015584
16	0.673382
17	0.649206
23	0.039058
25	0.444015
29	0.093763
38	0.277071
73	0.239424



LogReg Probability for XGBoost False Negatives



Number of False Negatives (Actual == 1, XGB Predicted == 0)

