brass

CUSTOMER CHURN PREDICTION

by Stanley Azuakola

brass

Industry Customers Products Company

Resources Login Open an account

ABOUT US

Making economic prosperity more accessible to everyone

There aren't enough successes, and growth is still almost impossible in most parts of Africa. That should change.

Brass started with a simple idea of building truly useful services that can enable commerce and economic prosperity for African local businesses.

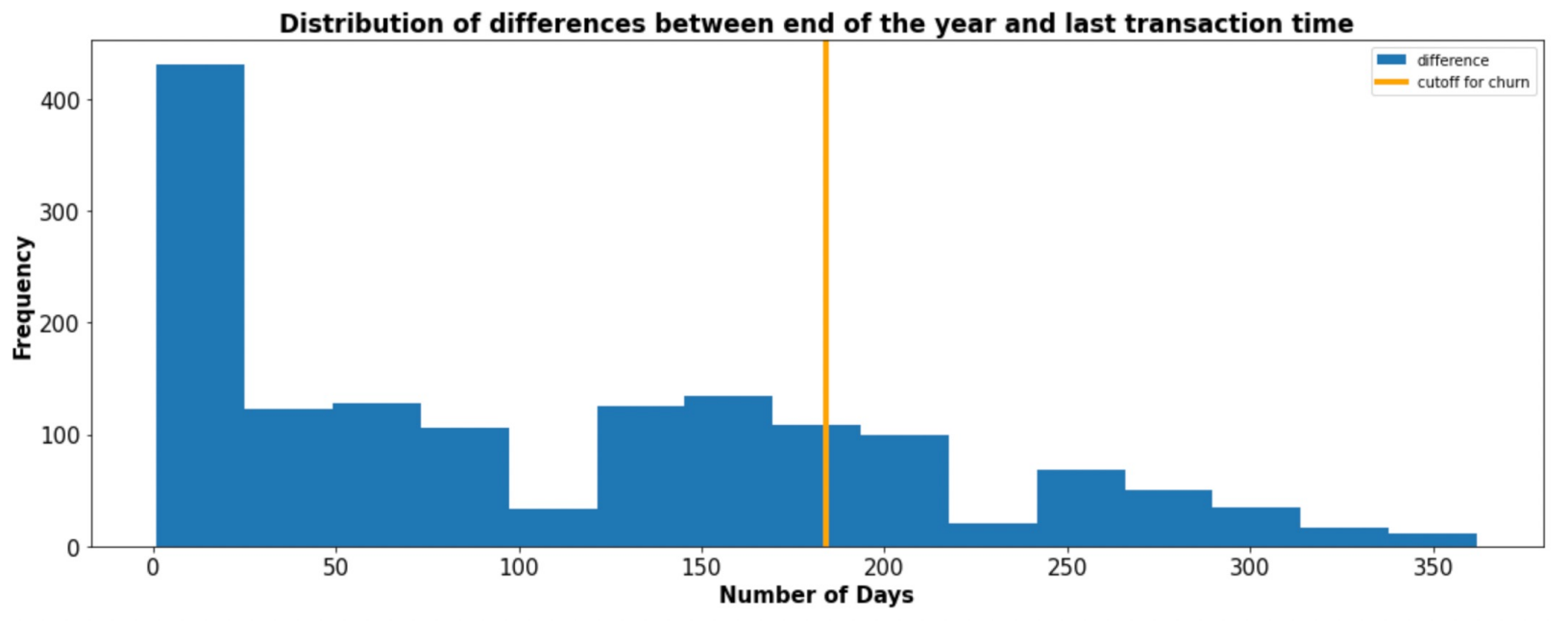
We aspire to do this by building fine products that can help boost business activity, remove inefficiencies and provide better performance and growth.

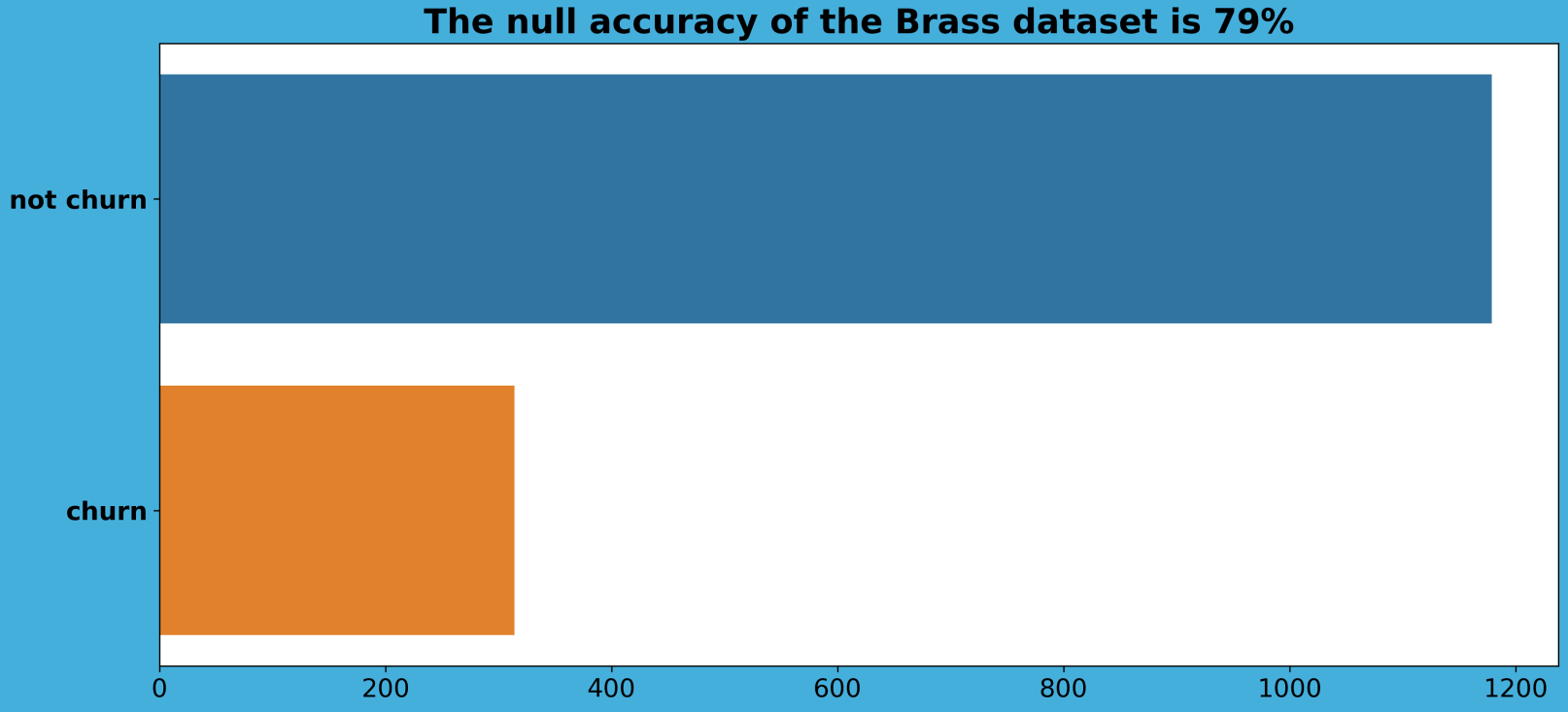
brass

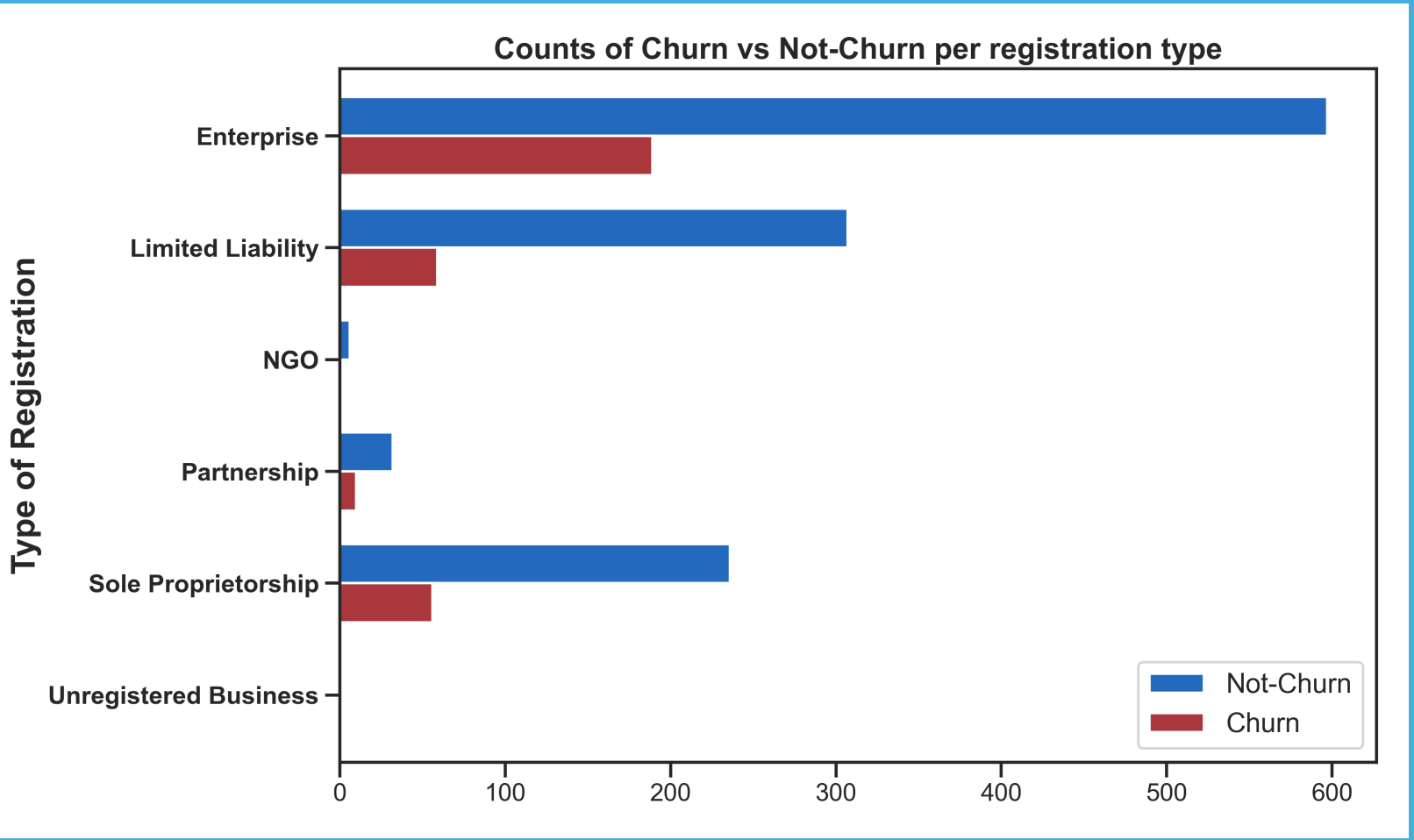
PROBLEM STATEMENT

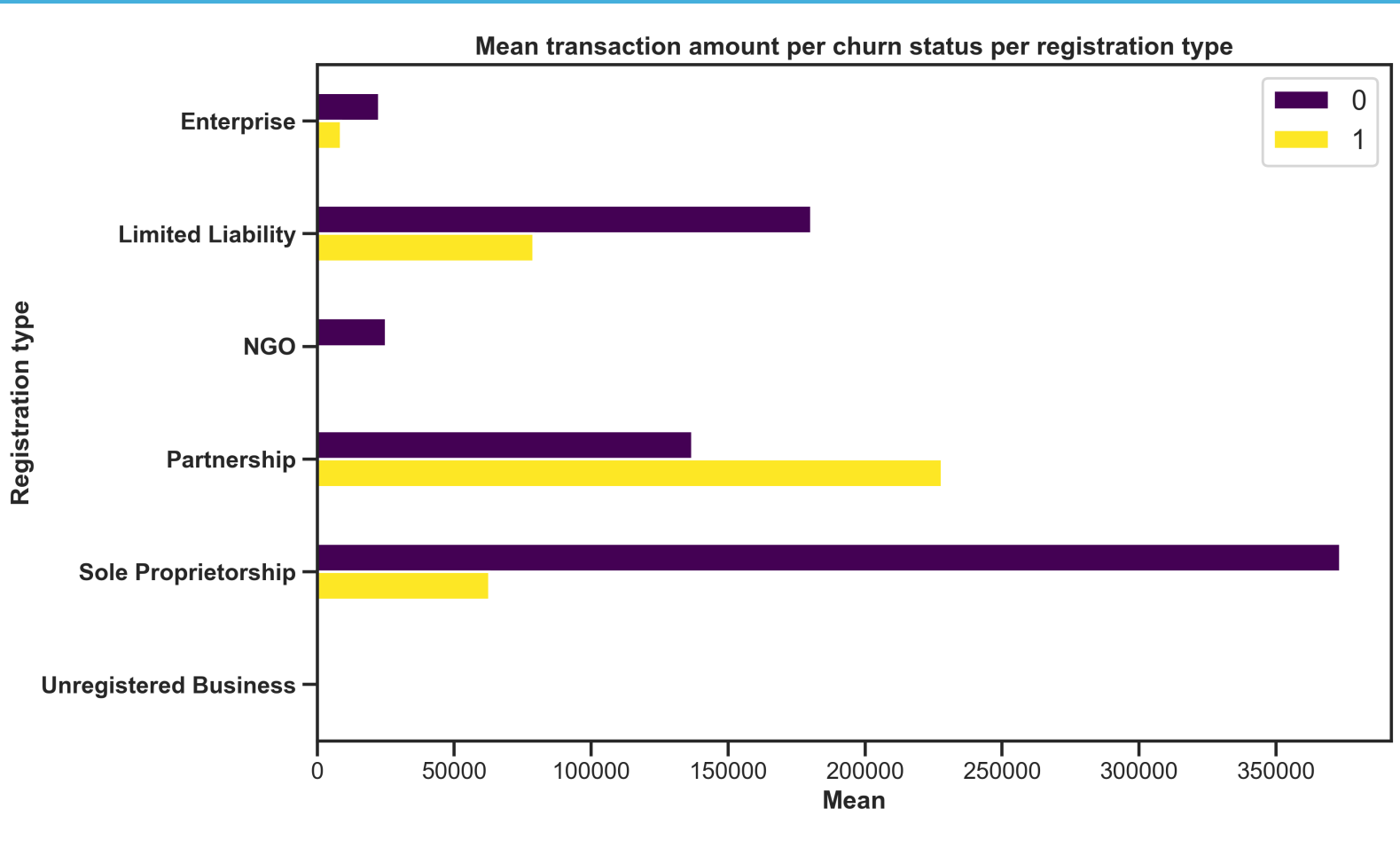
- Train a classifier to predict churn for Brass
- Define churn – **“we currently have no definition”**
- Engineer features - **“we can’t share for security reasons”**

CUSTOMER ACCOUNTS IN brass CAN STAY DORMANT FOR LONG PERIODS


Distribution of differences between end of the year and last transaction time
difference
cutoff for churn
Frequency
0
100
200
300
400
0
50
100
150
200
250
300
350
Number of Days

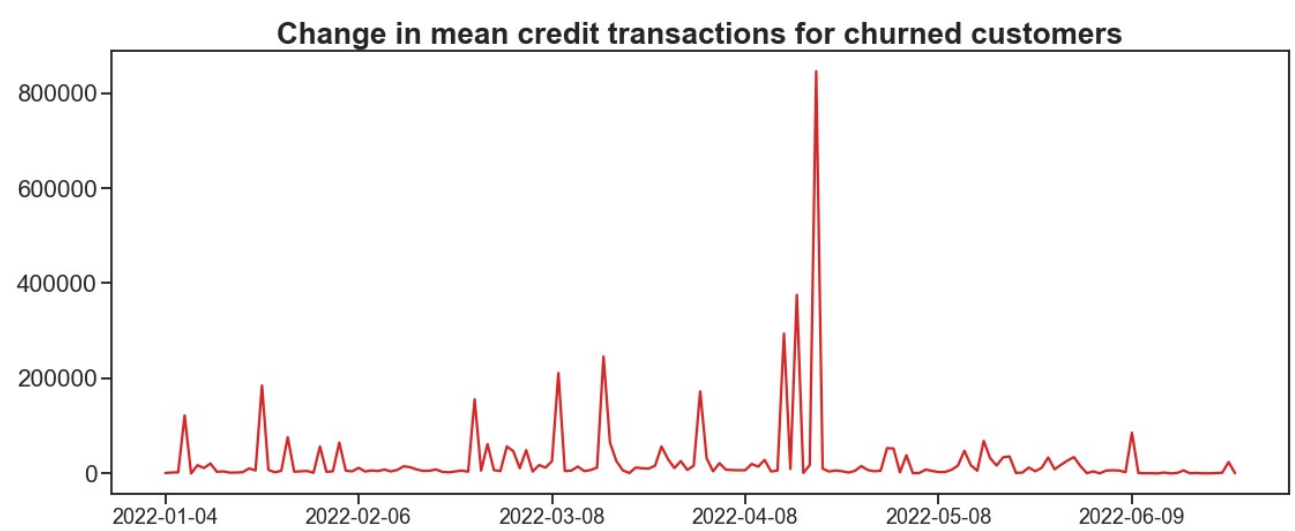

The null accuracy of the Brass dataset is 79%
not churn
churn
0
200
400
600
800
1000
1200

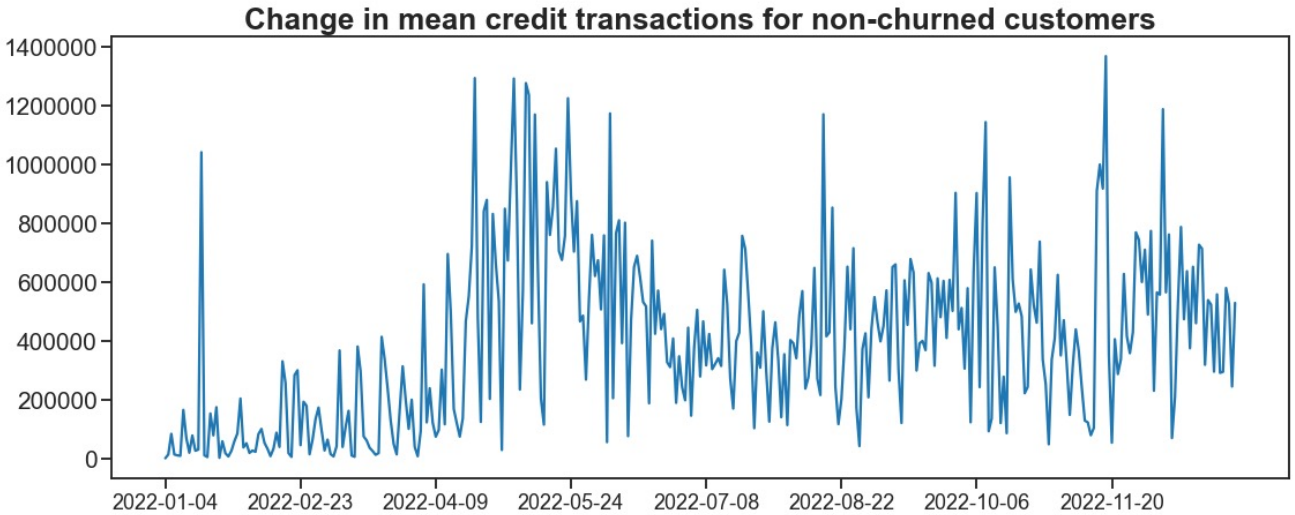

Counts of Churn vs Not-Churn per registration type
Type of Registration
Enterprise
Limited Liability
NGO
Partnership
Sole Proprietorship
Unregistered Business
0
100
200
300
400
500
600
Not-Churn
Churn

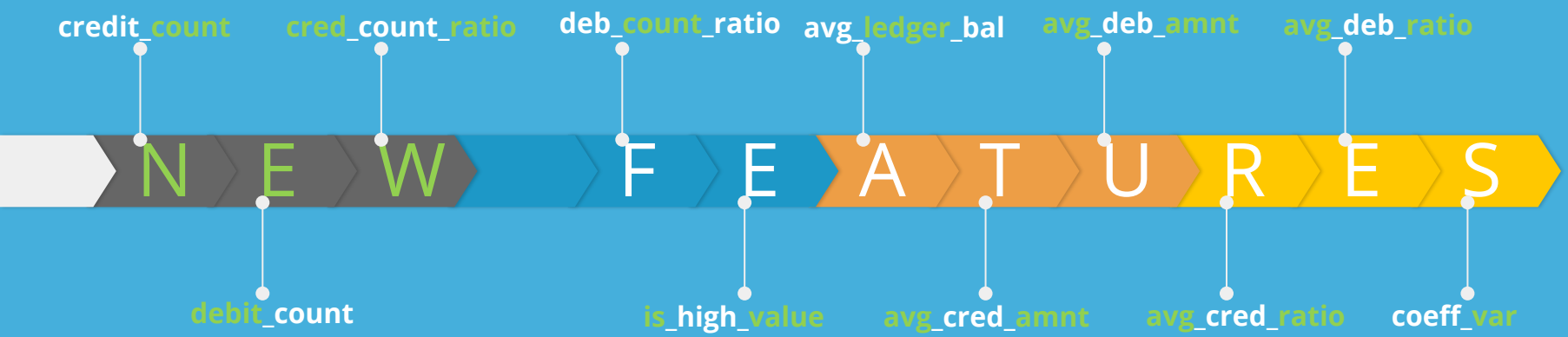

Mean transaction amount per churn status per registration type
0
1
Enterprise
Limited Liability
NGO
Partnership
Sole Proprietorship
Unregistered Business
Registration type
0
50000
100000
150000
200000
250000
300000
350000
Mean

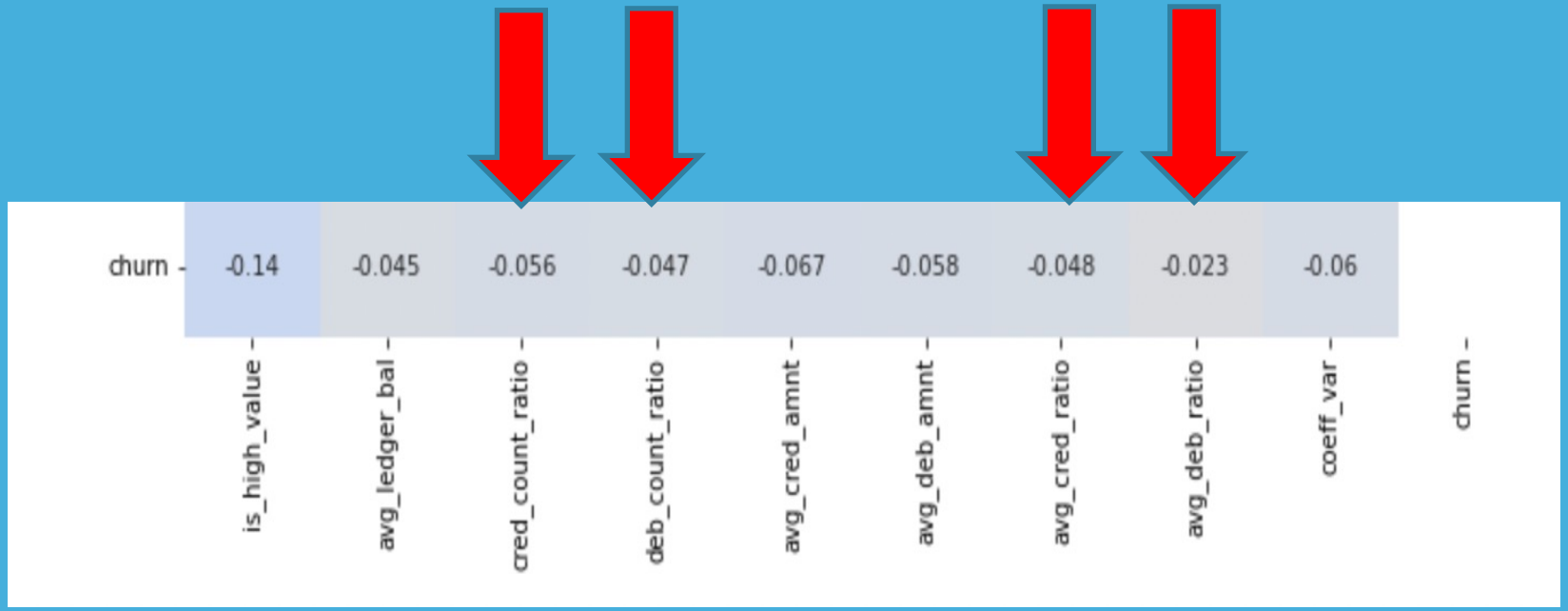
TIME FOR SOME FEATURE ENGINEERING

But wait... ☹


Change in mean credit transactions for churned customers
800000
600000
400000
200000
0
2022-01-04
2022-02-06
2022-03-08
2022-04-08
2022-05-08
2022-06-09


Change in mean credit transactions for non-churned customers
1400000
1200000
1000000
800000
600000
400000
200000
0
2022-01-04
2022-02-23
2022-04-09
2022-05-24
2022-07-08
2022-08-22
2022-10-06
2022-11-20


credit_count
cred_count_ratio
deb_count_ratio
avg_ledger_bal
avg_deb_amnt
avg_deb_ratio
NEW FEATURES
debit_count
is_high_value
avg_cred_amnt
avg_cred_ratio
coeff_var


churn
-0.14
-0.045
-0.056
-0.047
-0.067
-0.058
-0.048
-0.023
-0.06
is_high_value
avg_ledger_bal
cred_count_ratio
deb_count_ratio
avg_cred_amnt
avg_deb_amnt
avg_cred_ratio
avg_deb_ratio
coeff_var
churn

S/No	Feature	Description
1	`id`	Unique ID for each customer in the dataset
2	`registration_type`	Type of business entity based on registration of the customer
3	`industry`	Industry type for customer's business
4	`is_high_value`	a feature representing whether a customer is high or low value defined according to mean credit transactions
5	`avg_ledger_bal`	the average ledger balance per customer
6	`cred_count_ratio`	the count of `credit` transactions per number of days since registration per customer
7	`deb_count_ratio`	the count of `debit` transactions per number of days since registration per customer
8	`avg_cred_amnt`	the average credit amount per customer
9	`avg_deb_amnt`	the average credit amount per customer
10	`avg_cred_ratio`	the average credit amount per customer per number of days since registration
11	`avg_deb_ratio`	the average debit amount per customer per number of days since registration
12	`coeff_var`	the coefficient of variation of the `available_balance` for each customer
13	`churn`	did the customer churn or not - takes on a value of 0 for `not churn` and 1 for `churn`

FIVE EVALUATION METRICS

ACCURACY

- Customary
- Can be misleading

BALANCED ACCURACY

- Better!
- Average accuracy of + and -

RECALL

- Priority
- Say NO to false negatives

GEOMETRIC MEAN

- All about the balance

FBETA SCORE

- Like the F1 score but ‘BETA’

TIME FOR SOME MODELING

A LOGISTIC REGRESSION LOVE STORY

NULL ACCURACY

79%

LOGISTIC REGRESSION ACCURACY

79.7%

BALANCED ACCURACY

53.5%

RECALL

3%

Logistic Regression	Accuracy	Balanced Accuracy	Recall	Geometric Mean	Fbeta Score
Without dummies	79.7	51.6	3.2	43.3	74.2
With dummies	79.7	51.6	3.2	43.3	74.2

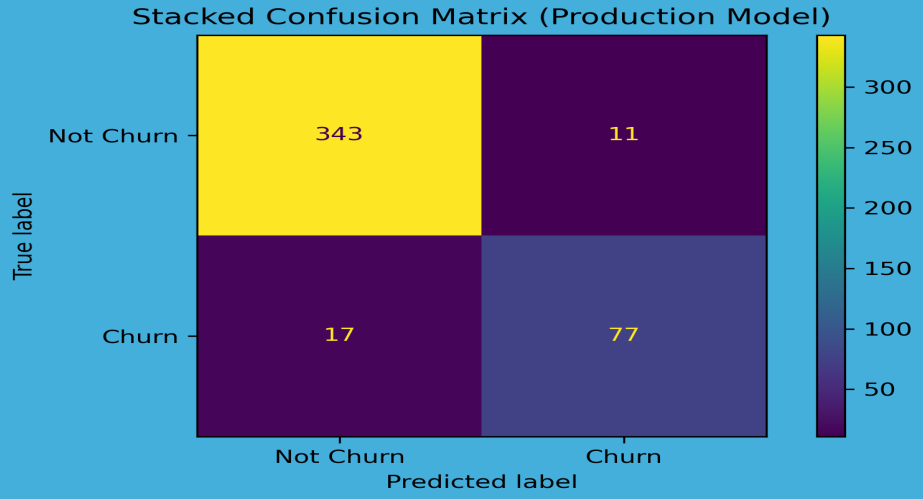
Decision Tree	Accuracy	Balanced Accuracy	Recall	Geometric Mean	Fbeta Score
Without dummies	87.9	80.2	67.0	79.9	87.8
With dummies	87.9	80.2	67.0	79.9	87.8

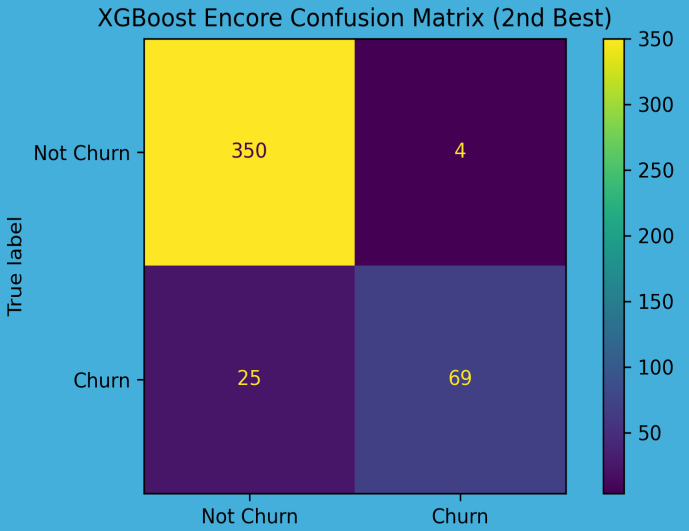
Random Forest	Accuracy	Balanced Accuracy	Recall	Geometric Mean	Fbeta Score
Without dummies	92.4	83.4	69.1	83.4	92.1
With dummies	91.5	82.1	65.6	81.6	91.2

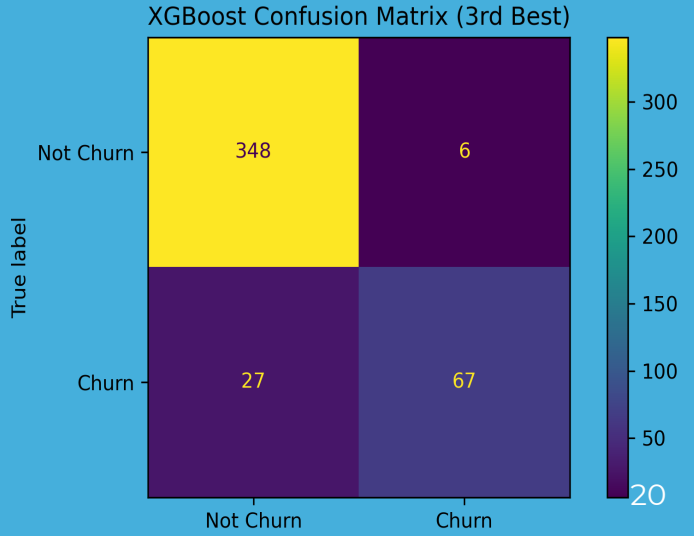
You know what, life is too short to stick with dummies . Move on without them.

- **Unknown**

1	**ALL OUR MODELS (MINUS THE DUMMY MODELS)**					
2		test_accuracy	test_bal_accuracy	test_recall	test_geom_mean	test_fbeta
3	Logistic Regression	0.796875	0.515957447	0.031915	0.432778714	0.741797
4	Random Forest	0.92410714	0.838682534	0.691489	0.834320699	0.920558
5	Decision Tree	0.87946429	0.802620507	0.670213	0.798933484	0.878144
6	Gradient Boost	0.91517857	0.821312658	0.659574	0.815931169	0.910886
7	XGBoost	0.92633929	0.847908402	**0.712766**	0.844273211	0.923454
8	Bagging Classifier	0.92410714	0.842589254	0.702128	0.838636682	0.920964
9	Stacked Classifier	0.92633929	0.847908402	**0.712766**	0.844273211	0.923454
10	RF Encore	0.91071429	0.81067436	0.638298	0.804478049	0.905772
11	Bagging Encore	0.90625	0.807849501	0.638298	0.801834246	0.901641
12	XGBoost Encore	0.93526786	0.861371559	**0.734043**	0.858195957	0.932632
13	Stacked Encore	0.9375	0.894037745	**0.819149**	0.892980695	0.936971

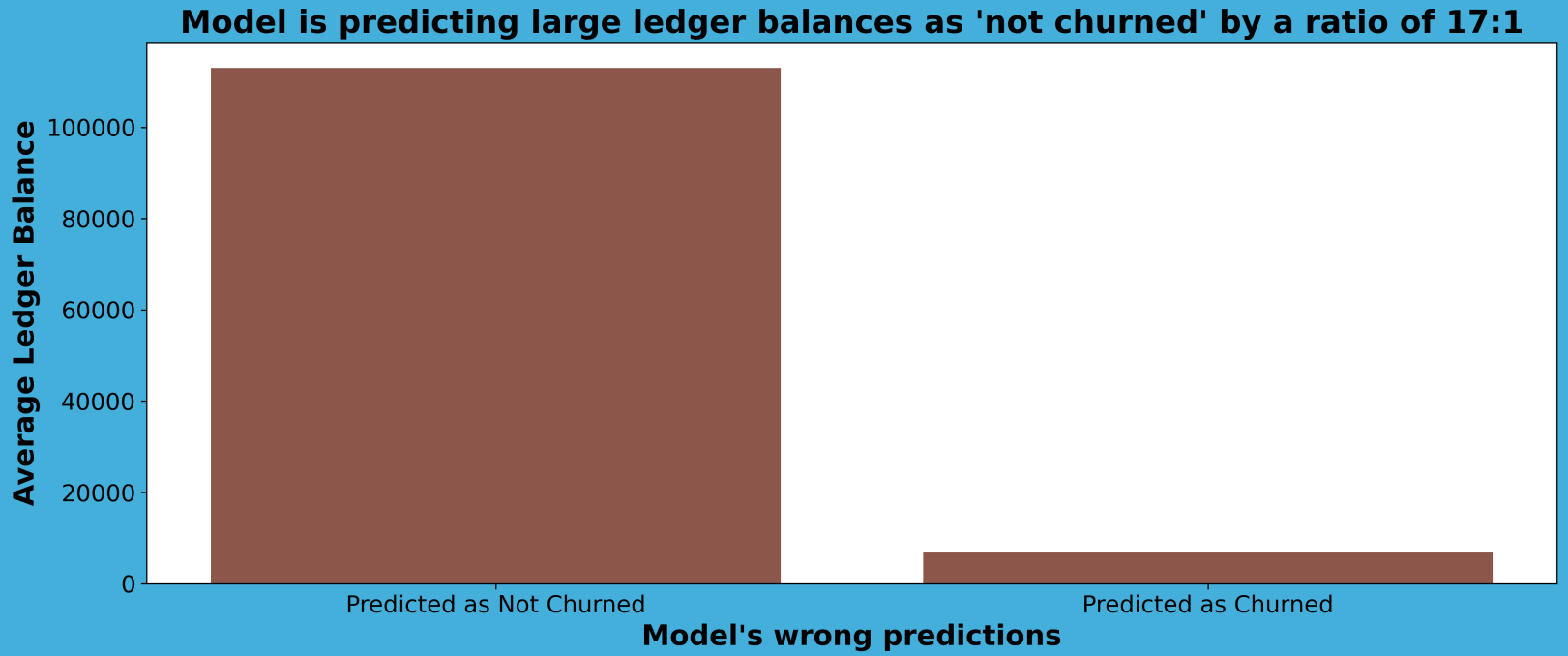

Stacked Confusion Matrix (Production Model)
Not Churn
Churn
True label
343
11
17
77
Not Churn
Churn
Predicted label
300
250
200
150
100
50


XGBoost Encore Confusion Matrix (2nd Best)
Not Churn
Churn
True label
350
4
25
69
Not Churn
Churn
350
300
250
200
150
100
50


XGBoost Confusion Matrix (3rd Best)
Not Churn
Churn
True label
348
6
27
67
Not Churn
Churn
300
250
200
150
100
50

	WHAT OUR MODEL GETS RIGHT
RECALL	82.0%: almost 10% more than next best score
BALANCED ACCURACY	89%: 3% more than next best score
GEOMETRIC MEAN	89%: 2% more than next best score
ACCURACY	94%: 15% more accurate than the null model
FALSE NEGATIVES	17 FNs: 32% improvement on next best model

WHAT OUR MODEL DOESN'T GET RIGHT



It appears that our model is sensitive to outliers

	id	registration_type	industry	transaction_type	amount	ledger_balance	available_balance	date	difference	churn
136183	12139	Enterprise	Logistics	credit	2050.00	0.00	0.00	2022-04-12	264	1
136184	12139	Enterprise	Logistics	credit	2500000.00	2050.00	2050.00	2022-04-16	260	1
136185	12139	Enterprise	Logistics	debit	2500000.00	2502050.00	2050.00	2022-05-20	226	1
136186	12139	Enterprise	Logistics	debit	53.75	2502050.00	2050.00	2022-05-20	226	1
136187	12139	Enterprise	Logistics	debit	1625.00	1996.25	1996.25	2022-06-01	214	1
136188	12139	Enterprise	Logistics	debit	1625.00	1996.25	1996.25	2022-06-10	205	1
136189	12139	Enterprise	Logistics	credit	1625.00	371.25	371.25	2022-06-10	205	1

$$\frac{2.5m}{3} = 800+ k$$

The fringe 'not churn' cases

	id	registration_type	industry	transaction_type	amount	ledger_balance	available_balance	date	difference	churn
55780	6657	Enterprise	General Services	credit	180.00	0.00	0.00	2022-02-10	325	0
55781	6657	Enterprise	General Services	debit	1000.00	2180.00	1180.00	2022-02-11	324	0
55782	6657	Enterprise	General Services	debit	10.75	2180.00	1180.00	2022-02-11	324	0
55783	6657	Enterprise	General Services	credit	2000.00	180.00	180.00	2022-02-11	324	0
55784	6657	Enterprise	General Services	debit	1000.00	1169.25	169.25	2022-06-10	205	0
55785	6657	Enterprise	General Services	debit	10.75	1169.25	169.25	2022-06-10	205	0
55786	6657	Enterprise	General Services	debit	1.00	158.50	158.50	2022-07-01	184	0

	id	registration_type	industry	transaction_type	amount	ledger_balance	available_balance	date	difference	churn
103629	10034	Limited Liability	Agriculture	debit	12000.00	17000.00	17000.00	2022-04-03	273	0
103630	10034	Limited Liability	Agriculture	credit	17000.00	0.00	0.00	2022-04-03	273	0
103631	10034	Limited Liability	Agriculture	debit	5000.00	5000.00	0.00	2022-06-14	201	0
103632	10034	Limited Liability	Agriculture	debit	4970.00	5000.00	30.00	2022-06-14	201	0
103633	10034	Limited Liability	Agriculture	debit	10.75	5000.00	30.00	2022-06-14	201	0
103634	10034	Limited Liability	Agriculture	credit	5000.00	0.00	0.00	2022-06-14	201	0
103635	10034	Limited Liability	Agriculture	debit	4.00	19.25	19.25	2022-07-01	184	0

brass

CONCLUSIONS

The problem:

Define churn, engineer features and predict churn for Brass online banking business.

The solution:

- **Defined churn using 75th percentile of time between most recent transactions and year end.**
- **Engineered 11 new features, using 9 of them for modeling.**
- **Trained classifier which predicted churn with 89% balanced accuracy, 82% recall, 89% geometric mean, 94% Fbeta score, and 94% accuracy.**

THANKS!

Any questions?