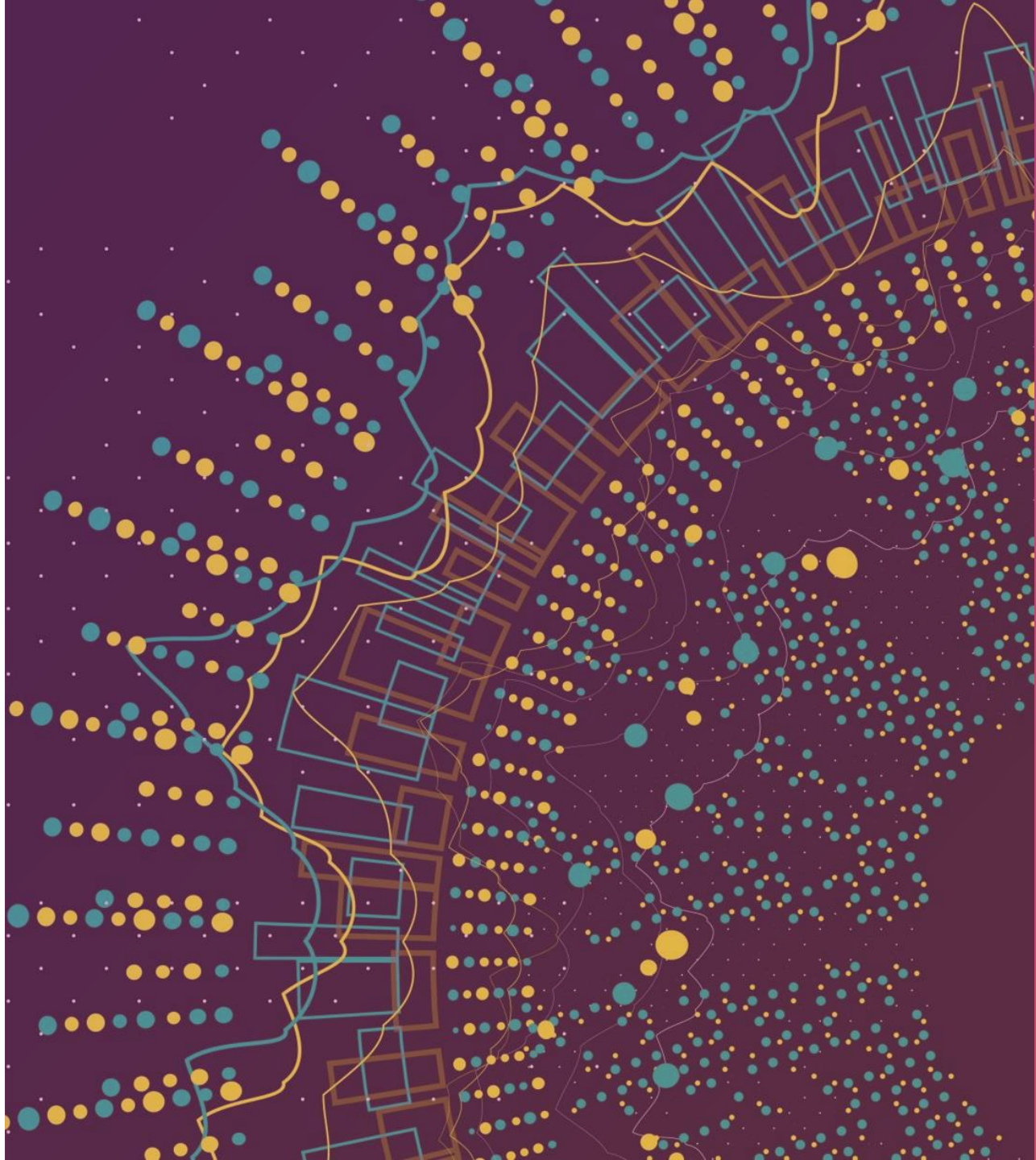
Ecommerce Marketing Strategy

Modeling website visitors' propensity to purchase



Business Context

- **Problem:** This ecommerce company is currently marketing to all website visitors equally even though only a small percentage of visitors are likely to actually make a purchase
- **Questions:**
  1. What variables/activity make a customer more likely to place an order?
  2. What customers should the ecommerce company market to?
  3. At what stage of the customer journey should the ecommerce company implement a marketing campaign?

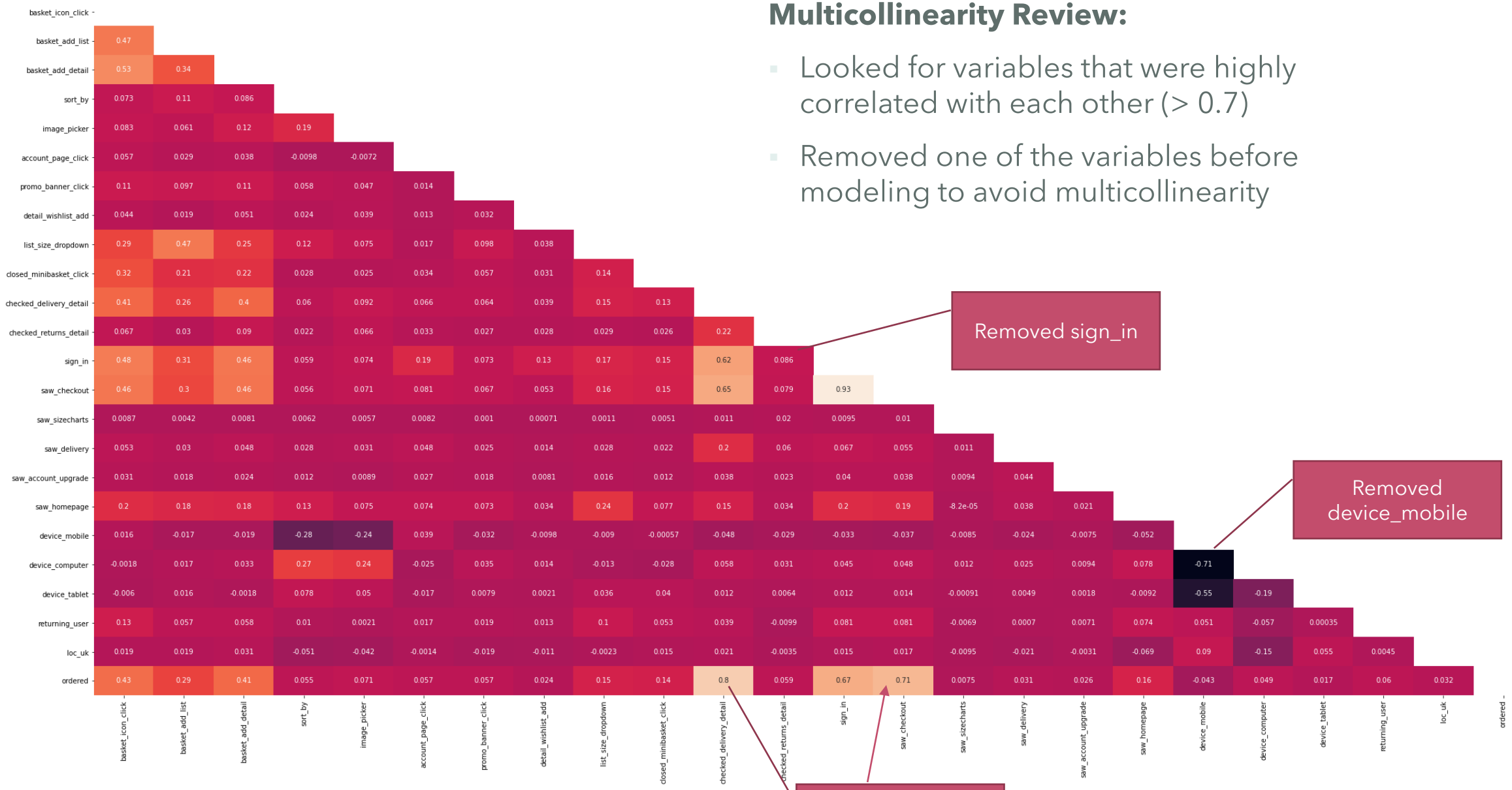
Data

- Website traffic for one day
- 455,401 customers
- 25 variables

Variable	Definition
UserID	A unique identifier for the visitor
basket_icon_click	Did the visitor click on the shopping basket icon?
basket_add_list	Did the visitor add a product to their shopping cart on the 'list' page?
basket_add_detail	Did the visitor add a product to their shopping cart on the 'detail' page?
sort_by	Did the visitor sort products on a page?
image_picker	Did the visitor use the image picker?
account_page_click	Did the visitor visit their account page?
promo_banner_click	Did the visitor click on a promo banner?
detail_wishlist_add	Did the visitor add a product to their wishlist from the 'detail' page?
list_size_dropdown	Did the visitor interact with a product dropdown?
closed_minibasket_click	Did the visitor close their mini shopping basket?
checked_delivery_detail	Did the visitor view the delivery FAQ area on a product page?
checked_returns_detail	Did the visitor check the returns FAQ area on a product page?
sign_in	Did the visitor sign in to the website?
saw_checkout	Did the visitor view the checkout?
saw_sizecharts	Did the visitor view a product size chart?
saw_delivery	Did the visitor view the delivery FAQ page?
saw_account_upgrade	Did the visitor view the account upgrade page?
saw_homepage	Did the visitor view the website homepage?
device_mobile	Was the visitor on a mobile device?
device_computer	Was the visitor on a desktop device?
device_tablet	Was the visitor on a table device?
returning_user	Was the visitor new or returning?
loc_uk	Was the visitor located in the UK, based on their IP address?
ordered	Did the customer place an order?

Multicollinearity Review:

- Looked for variables that were highly correlated with each other (> 0.7)
- Removed one of the variables before modeling to avoid multicollinearity



```
1 train_df.corr()['ordered']
```

```
basket_icon_click           0.428334
basket_add_list             0.287666
basket_add_detail           0.414420
sort_by                     0.054636
image_picker                0.071492
account_page_click          0.057279
promo_banner_click          0.056533
detail_wishlist_add         0.023516
list_size_dropdown          0.154867
closed_minibasket_click     0.140011
checked_delivery_detail     0.798720
checked_returns_detail      0.059484
sign_in                     0.665556
saw_checkout                0.708986
saw_sizecharts              0.007548
saw_delivery                0.031461
saw_account_upgrade         0.025857
saw_homepage                0.157778
device_mobile              -0.042907
device_computer             0.049208
device_tablet               0.016939
returning_user              0.060295
loc_uk                      0.031643
ordered                     1.000000
Name: ordered, dtype: float64
```

Highest Correlations with Ordered

Checked_delivery_detail = 0.798

Saw_checkout = 0.708

Negative Correlation with Ordered

Device_mobile = -0.043

-Perhaps people are using their mobile device to browse instead of order

-Maybe we don't target this segment for the marketing campaign

Imbalanced Classifications Issue:

- This occurs when the dataset is highly skewed in the class distribution.
- Of the 455,401 unique customers, 436,308 did not place an order (Ordered =0)
- This bias in the training dataset can influence many machine learning algorithms, leading some to ignore the minority class entirely
- Techniques to counter include:

  -Over sample – duplicate random records from the minority class

  -Under sample- remove random records from the majority class

  -Use the "Recall" calculation for accuracy score- total number of true positive predictions divided by the sum of the true positives and the false negatives
- For this project, the recall calculation is used for the accuracy score

Model 1

- Model 1 ignored the independent variables “sign_in” and “device_mobile” since they are correlated with another variable
- Used scikit-learn Python package to:
  - Perform a random 80-20 train-test split on the data
  - Fit a logistic regression model on the train data.

Logit Regression Results

Dep. Variable:	ordered	No. Observations:	455401
Model:	Logit	Df Residuals:	455379
Method:	MLE	Df Model:	21
Date:	Tue, 07 Dec 2021	Pseudo R-squ.:	0.8726
Time:	15:48:50	Log-Likelihood:	-10095.
converged:	True	LL-Null:	-79247.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-11.7185	0.148	-79.085	0.000	-12.009	-11.428
basket_icon_click	0.4120	0.043	9.690	0.000	0.329	0.495
basket_add_list	0.4007	0.054	7.469	0.000	0.296	0.506
basket_add_detail	0.5723	0.043	13.405	0.000	0.489	0.656
sort_by	-0.1443	0.080	-1.805	0.071	-0.301	0.012
image_picker	0.0191	0.081	0.236	0.813	-0.140	0.178
account_page_click	-0.6021	0.108	-5.568	0.000	-0.814	-0.390
promo_banner_click	-0.1663	0.088	-1.896	0.058	-0.338	0.006
detail_wishlist_add	-1.0711	0.127	-8.434	0.000	-1.320	-0.822
list_size_dropdown	-0.1791	0.051	-3.538	0.000	-0.278	-0.080
closed_minibasket_click	0.1685	0.071	2.358	0.018	0.028	0.308
checked_delivery_detail	6.2726	0.075	83.415	0.000	6.125	6.420
checked_returns_detail	-1.0142	0.073	-13.895	0.000	-1.157	-0.871
sign_in	5.0545	0.099	51.073	0.000	4.860	5.248
saw_sizecharts	-0.1402	0.499	-0.281	0.779	-1.118	0.838
saw_delivery	-1.9901	0.082	-24.329	0.000	-2.150	-1.830
saw_account_upgrade	-0.6163	0.211	-2.919	0.004	-1.030	-0.202
saw_homepage	0.3704	0.039	9.382	0.000	0.293	0.448
device_computer	0.5260	0.052	10.146	0.000	0.424	0.628
device_tablet	0.4706	0.060	7.902	0.000	0.354	0.587
returning_user	0.4121	0.040	10.311	0.000	0.334	0.490
loc_uk	0.9763	0.084	11.571	0.000	0.811	1.142

Results

Confusion Matrix			
Predicted	Did not order	86715	547
	Ordered	42	3777
		Did not order	Ordered
		Actual	

- Hit Rate: (86715+3777)/(86715+3777+42+547) = 99.35%
- Precision: 3777/(3777+42) = 98.90%
- Recall: 3777/(3777+547) = **87.35%**

- **Imbalanced classification:** We needed to look at precision and recall because there were way more customers that did not place an order than customers who did place and order.
- **Recall** focuses on the model's ability to predict the positive class (Ordered)

Model 2 – Remove variables with p-value > 0.05

- Removed - "sort_by", "image_picker", promo_banner_click, saw_checkout & saw_sizecharts.

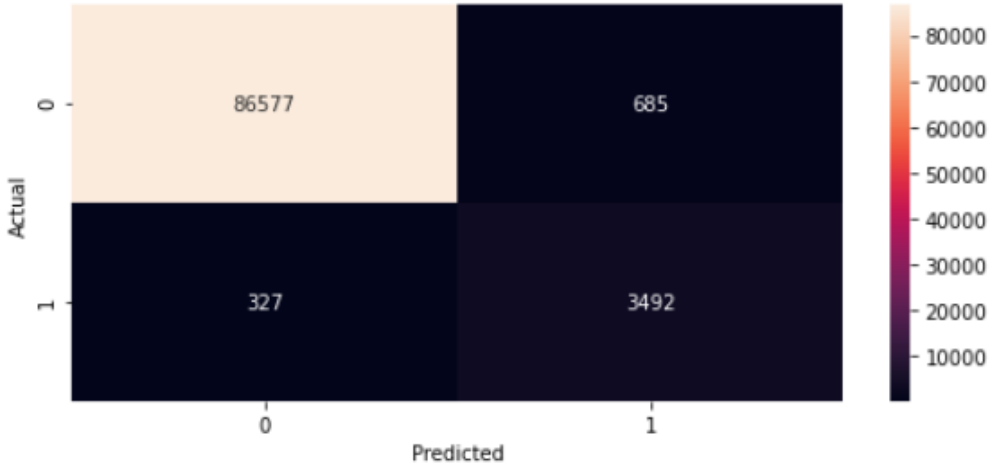
- Doing this caused account_page_click, closed_minibasket_click and _saw_account_upgrade to exceed our p-value threshold of 0.05, so these will be excluded from the calcualtion

Logit Regression Results

Dep. Variable:	ordered	No. Observations:	455401
Model:	Logit	Df Residuals:	455384
Method:	MLE	Df Model:	16
Date:	Tue, 07 Dec 2021	Pseudo R-squ.:	0.8264
Time:	17:50:29	Log-Likelihood:	-13757.
converged:	True	LL-Null:	-79247.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-10.2961	0.114	-90.552	0.000	-10.519	-10.073
basket_icon_click	1.1266	0.035	32.173	0.000	1.058	1.195
basket_add_list	1.1710	0.046	25.206	0.000	1.080	1.262
basket_add_detail	1.5859	0.033	48.084	0.000	1.521	1.651
account_page_click	-0.0350	0.122	-0.286	0.775	-0.274	0.204
detail_wishlist_add	-0.5315	0.138	-3.843	0.000	-0.803	-0.260
list_size_dropdown	-0.4017	0.041	-9.742	0.000	-0.483	-0.321
closed_minibasket_click	-0.0783	0.065	-1.204	0.229	-0.206	0.049
checked_delivery_detail	7.7915	0.074	105.280	0.000	7.646	7.937
checked_returns_detail	-1.7566	0.058	-30.364	0.000	-1.870	-1.643
saw_delivery	-2.5026	0.075	-33.209	0.000	-2.650	-2.355
saw_account_upgrade	-0.3848	0.211	-1.824	0.068	-0.798	0.029
saw_homepage	0.6025	0.033	18.526	0.000	0.539	0.666
device_computer	0.5085	0.037	13.619	0.000	0.435	0.582
device_tablet	0.5008	0.048	10.541	0.000	0.408	0.594
returning_user	0.7244	0.033	22.227	0.000	0.661	0.788
loc_uk	0.9934	0.077	12.914	0.000	0.843	1.144

Results



- Hit Rate: = 99.30%
- Recall: = 83.60%

- **Imbalanced classification:** We needed to look at precision and recall because there were way more customers that did not place an order than customers who did place and order.
- **Recall** focuses on the model's ability to predict the positive class (Ordered)

Model 3 – Narrowed to Five Variables

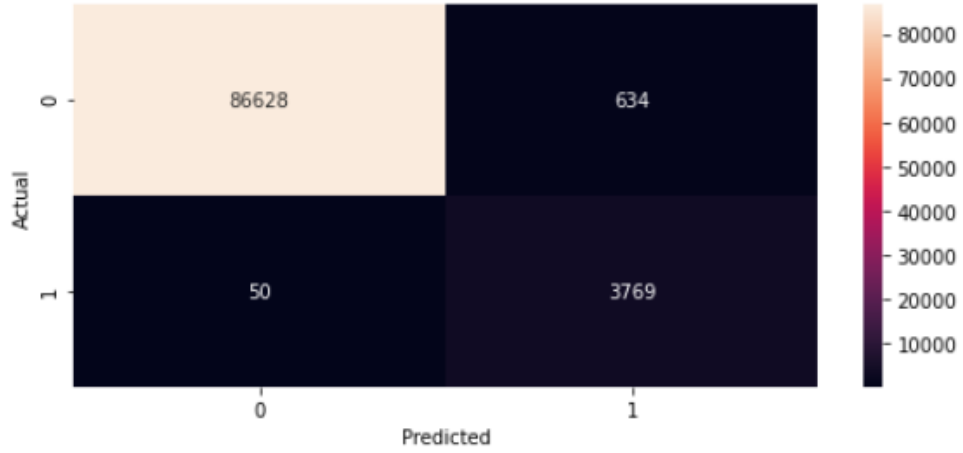
- Added sign_in back in

Logit Regression Results

Dep. Variable:	ordered	No. Observations:	455401
Model:	Logit	Df Residuals:	455395
Method:	MLE	Df Model:	5
Date:	Tue, 07 Dec 2021	Pseudo R-squ.:	0.8634
Time:	23:22:37	Log-Likelihood:	-10824.
converged:	True	LL-Null:	-79247.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-10.6648	0.117	-91.009	0.000	-10.895	-10.435
basket_icon_click	0.4589	0.038	12.180	0.000	0.385	0.533
basket_add_detail	0.4879	0.039	12.519	0.000	0.412	0.564
checked_delivery_detail	6.1516	0.074	82.659	0.000	6.006	6.298
sign_in	5.3575	0.097	55.102	0.000	5.167	5.548
returning_user	0.3906	0.038	10.304	0.000	0.316	0.465

Results



- Hit Rate: = 99.30%
- Recall: = 85.60%

- **Imbalanced classification:** We needed to look at precision and recall because there were way more customers that did not place an order than customers who did place and order.
- **Recall** focuses on the model's ability to predict the positive class (Ordered)

Recommendations

- This ecommerce giant should focus their marketing on visitors that:

1. Signed into their account 0.66 correlation to ordering
2. Visited the delivery FAQs section 0.79 correlation with ordering
3. Click on the shopping basket icon 0.42 correlation with ordering
4. Don’t bother to target the mobile device customers

Recommendations continued:

1. Tailor the website to make it easier for customer to check out:
2. Add “ready to checkout link” to the FAQ for delivery page
3. When customers sign in, show any items already in their cart

Potential Issues:

1. The dataset is web traffic for only one day.
2. Imbalance Issue
3. This is “point in time” data and not panel