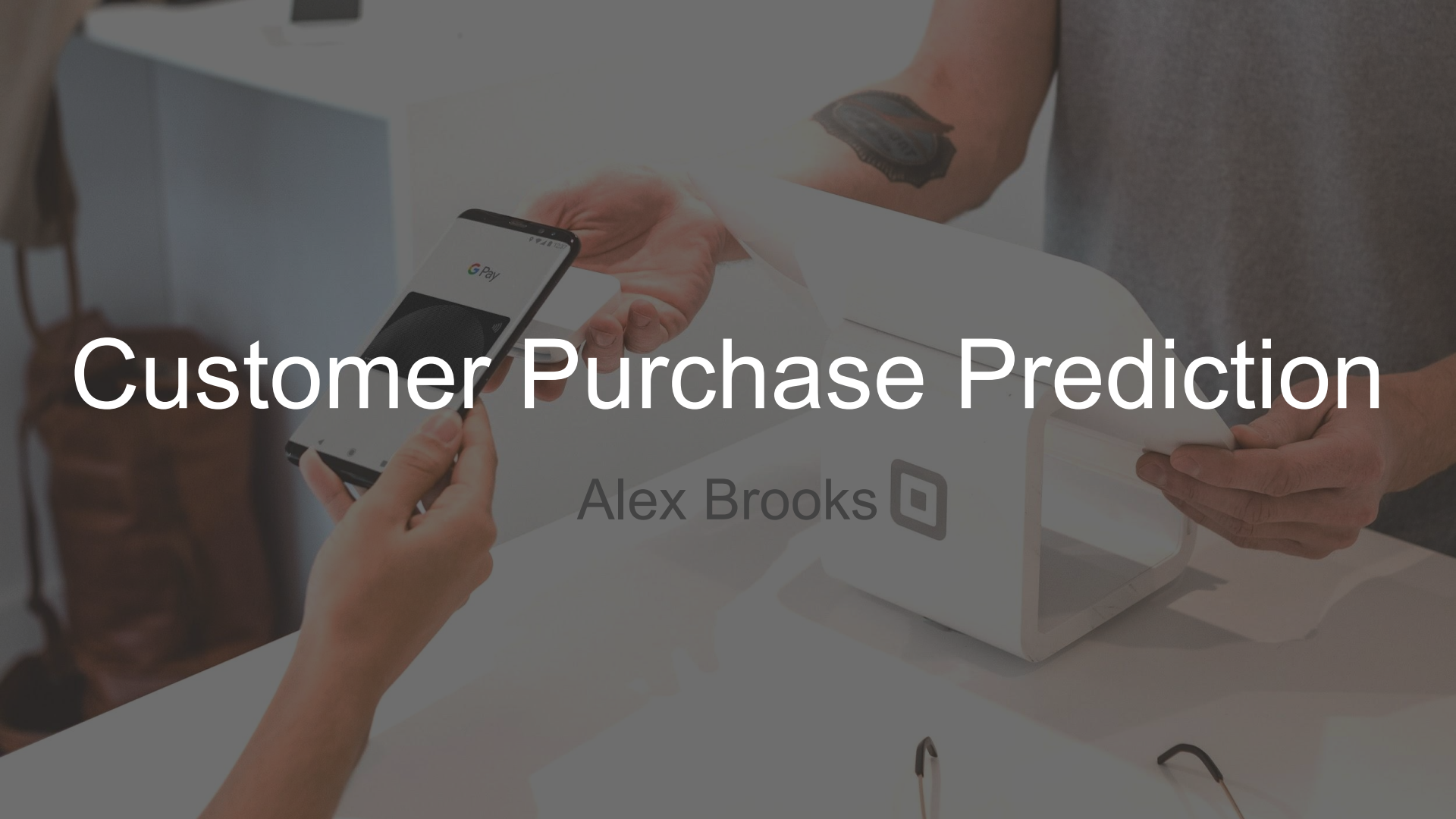
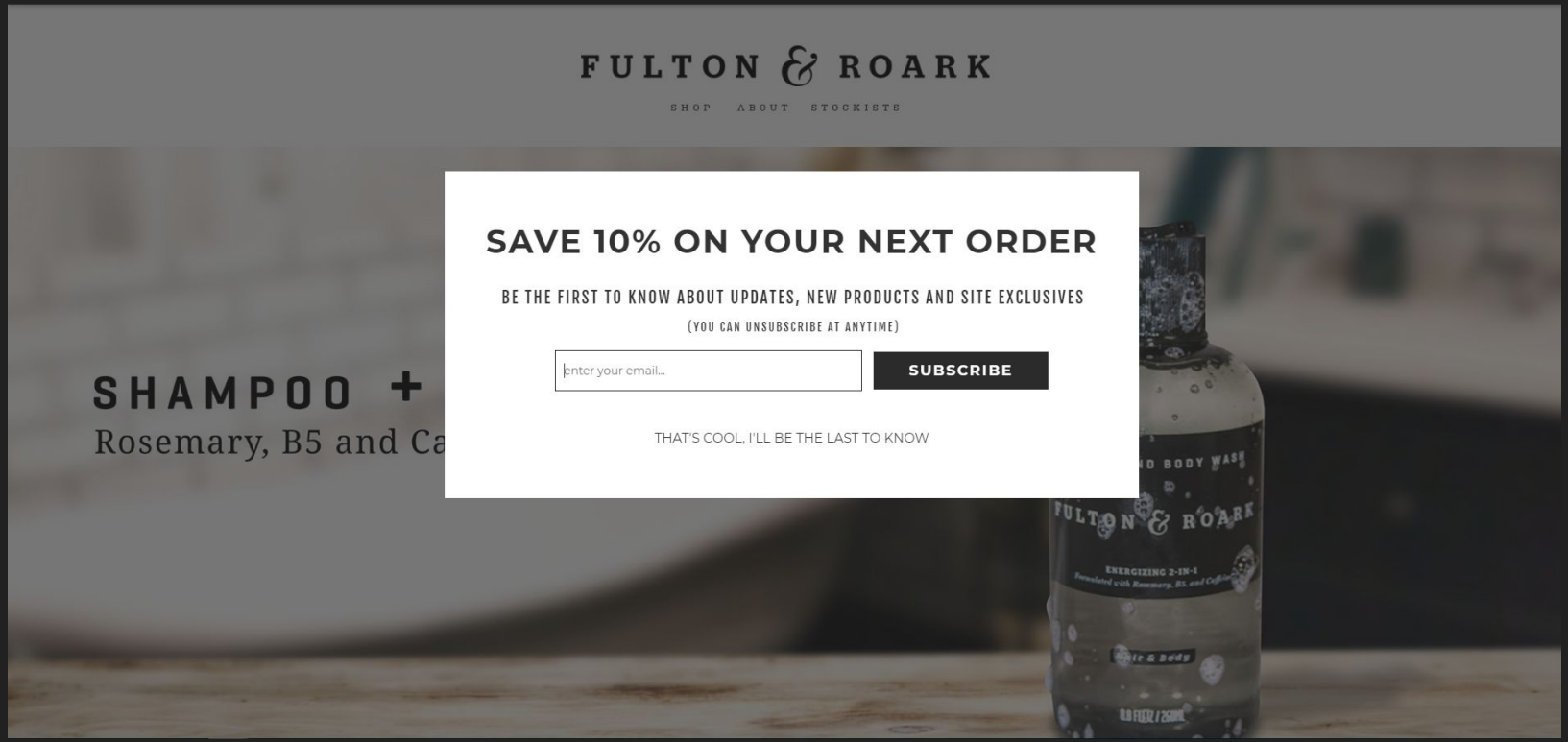


Customer Purchase Prediction

Alex Brooks

I think we've all seen this before



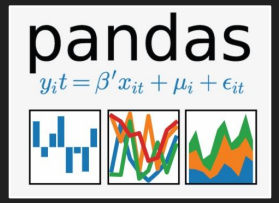
A fraction of those users will actually buy from you

event_time	event_type	product_id	category_id	category_code	brand	price	user_id	user_session
2019-10-01 00:00:26	view	5875317	2.029083e+18	None	None	1.59	474232307	445f2b74-5e4c-427e-b7fa-6e0a28b156fe
2019-10-01 00:00:28	view	5692917	1.487580e+18	None	lianail	5.54	555446068	4257671a-efc8-4e58-96c2-3ab457916d78
2019-10-01 00:00:48	view	5819638	1.487580e+18	None	None	21.75	546705258	3b5c65c0-bb1c-453b-b340-4ebf973a3136
2019-10-01 00:01:11	view	5733176	1.487580e+18	None	None	0.60	543446752	4a01cad9-7368-fd9e-d907-d4d85de0b55c
2019-10-01 00:01:16	view	5712497	1.487580e+18	None	f.o.x	6.03	555446068	4257671a-efc8-4e58-96c2-3ab457916d78
...	...	...	...	...	...	...	...	...
2019-11-30 23:59:24	view	5694628	1.487580e+18	None	yoko	3.65	576802932	f5b0e79b-0470-4256-ba47-4166ea4f05f5
2019-11-30 23:59:32	view	5795387	1.487580e+18	None	ingarden	7.14	576802932	2dc9ed07-93bb-47db-abe7-e9d88ed7ae94
2019-11-30 23:59:37	view	5699730	1.487580e+18	None	None	2.70	422196217	dd6d8240-0896-4965-9344-110648581a51
2019-11-30 23:59:46	view	5830317	1.487580e+18	None	None	4.76	457678989	ee50b160-a4db-4722-8751-6812c5b38295
2019-11-30 23:59:47	view	5733064	1.487580e+18	None	beautix	9.37	422196217	ab5e6dd5-8700-4ecc-a300-9f1eca5d1a95

Objective

Predict if a customer will make a purchase based on user action, not transactional data.

Workflow



Data Wrangling

- Load into Postgres
- Extract w/ SQL

Data Cleaning
(& Pre-Processing)

- Modify target
- Reduce data leakage

Feature Engineering

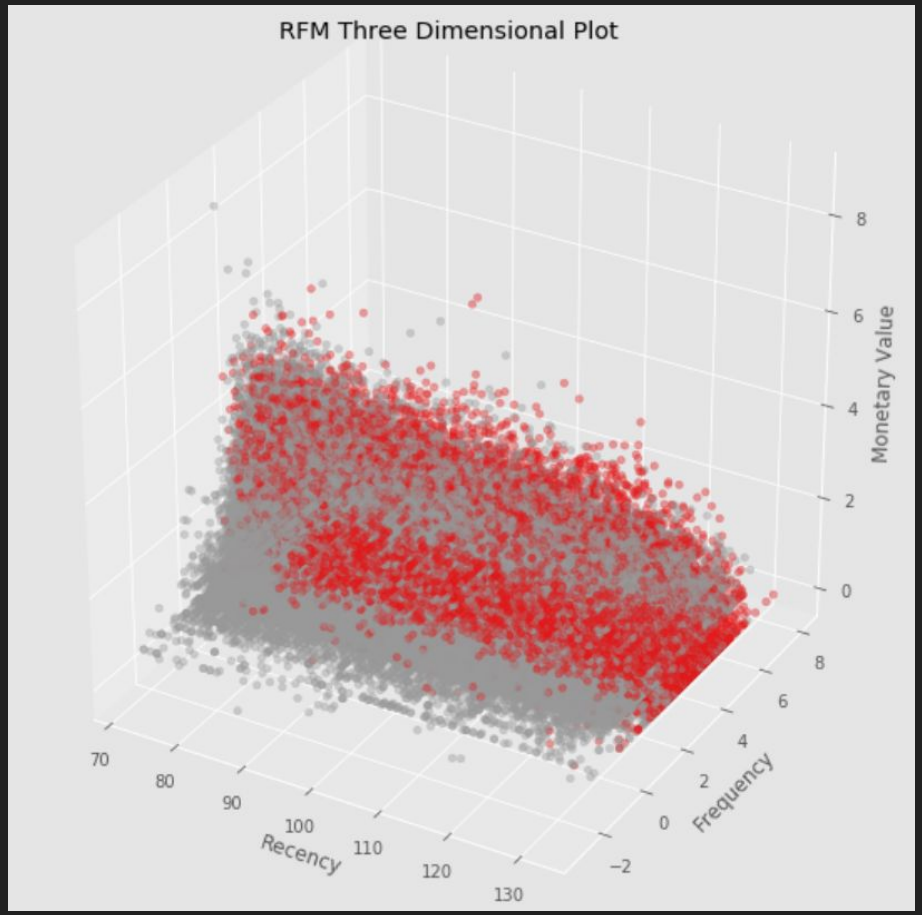
- Create features w/ SQL

Modeling

- Run models in AWS

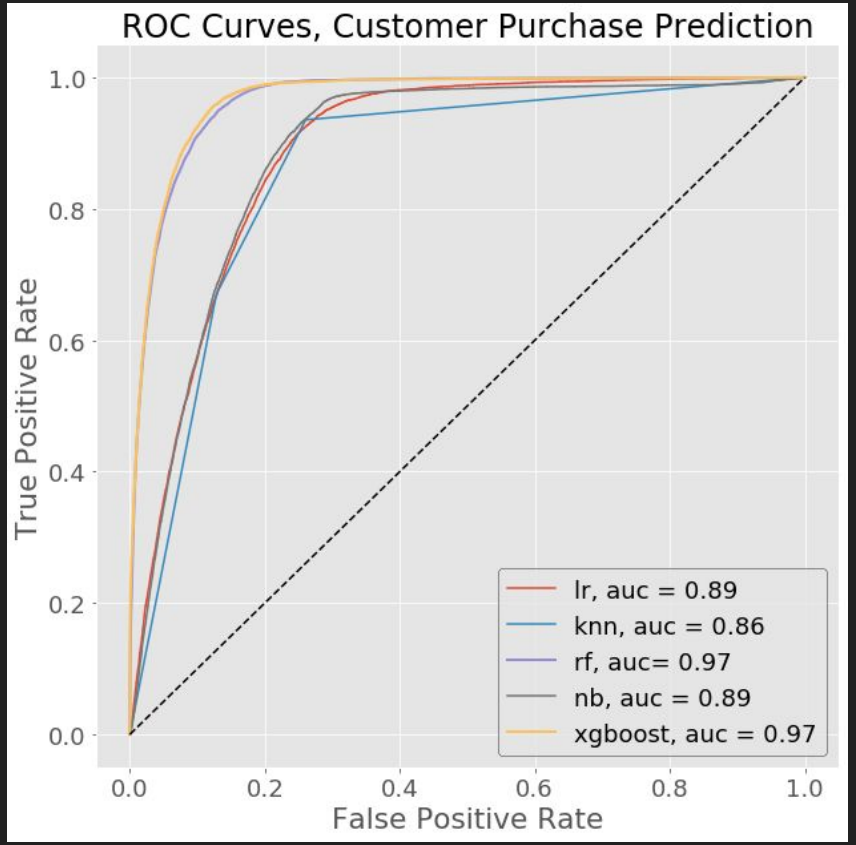
Feature Engineering

- Actions/rate of action
- Brand affinity
- Pricing attributes
- Recency, Frequency, Monetary Value (RFM)
  - Raw values + **Discretization**



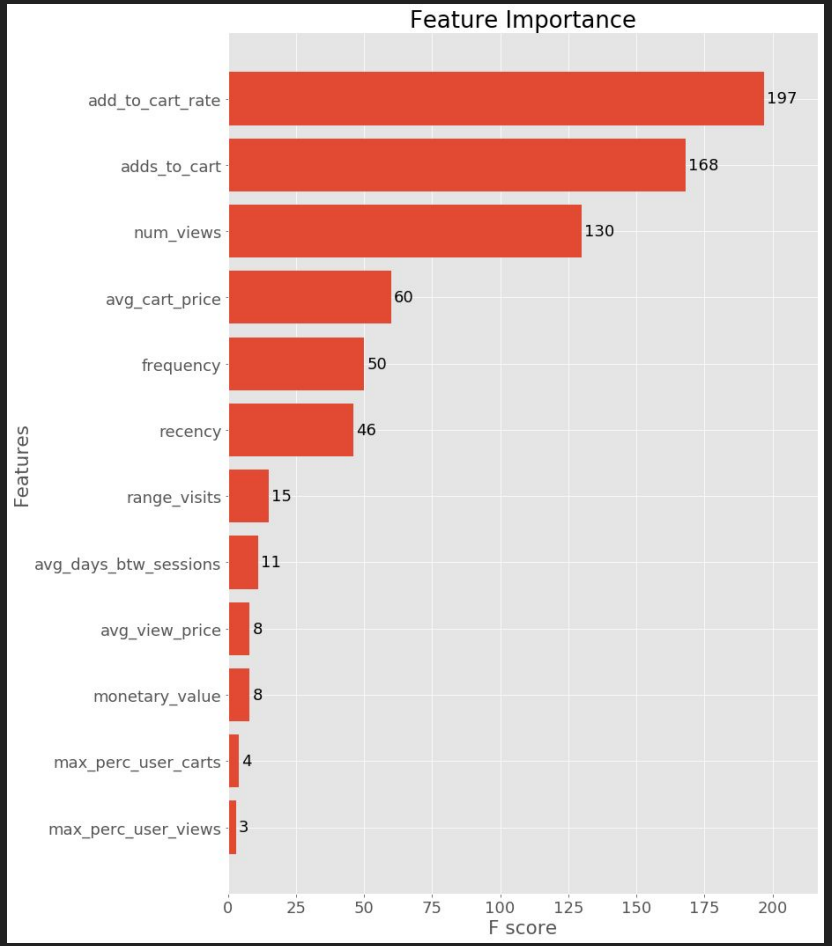
Model performance

- **XGBoost model**
  - 96% customer classification
  - F1 score of 0.79
- Right-skewed metrics led to overfitting
- Collinearity



Key Takeaways

- Action items:
  - Retention of prospective customers: incentives, cart reminders
  - Create targeted email list segments
  - Referral incentives



Future work

- Predict product purchases
- Time-based features
  - Session duration
  - Time on page
- Calculate F_beta using churn cost

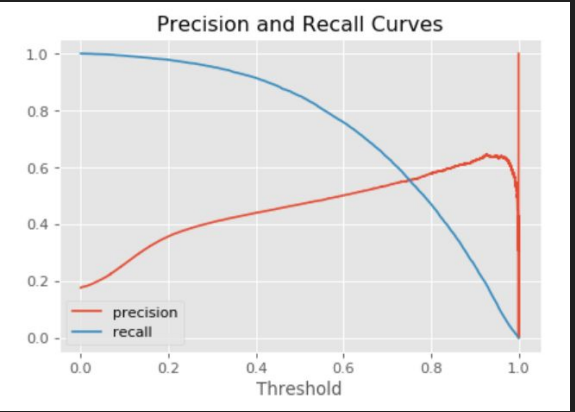
Questions?

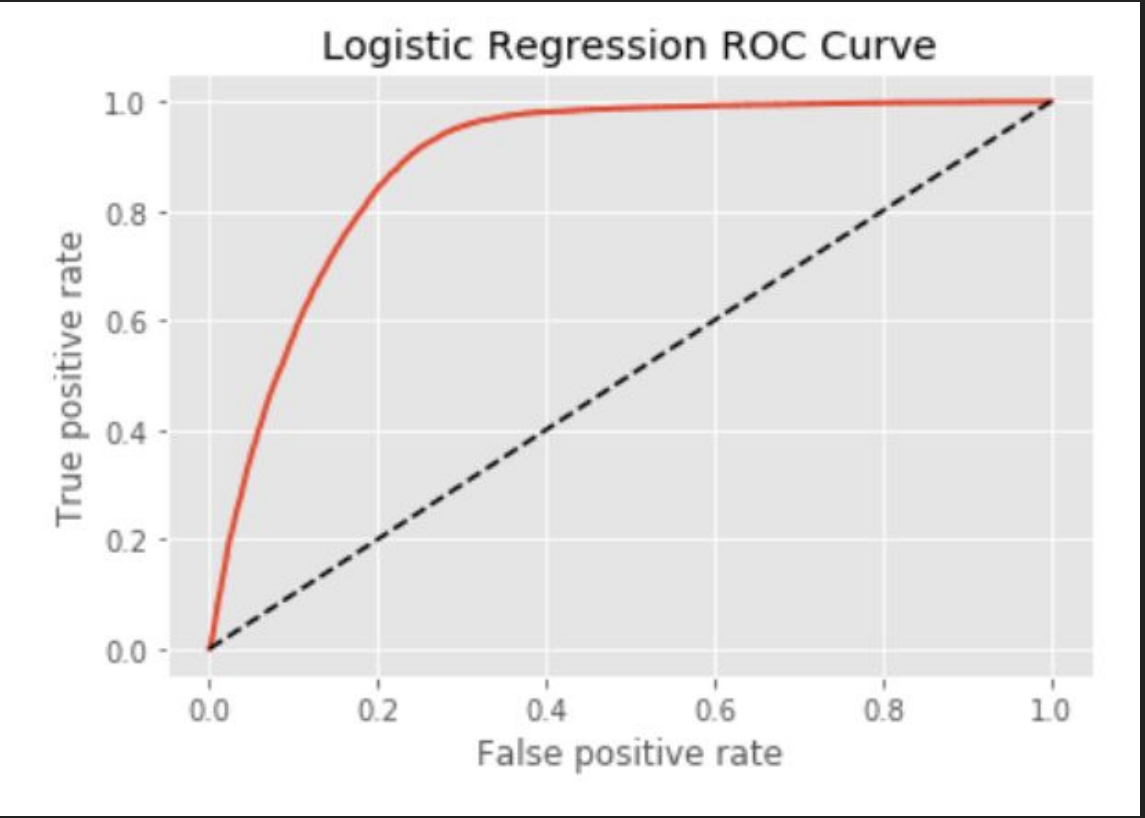
Appendix

Model Performance Overview

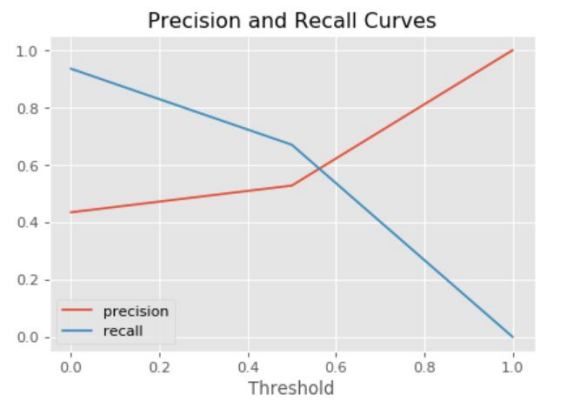
Model	Accuracy Score	ROC/AUC Score	F1 Score	Precision	Recall
Logistic Regression	0.80	0.89	0.61	0.47	0.85
K-Nearest Neighbors (k=2)	0.84	0.86	0.59	0.53	0.67
Random Forest	0.88	0.97	0.73	0.59	0.96
Naive Bayes	0.79	0.89	0.61	0.45	0.92
XGBoost	0.88	0.97	0.79	0.60	0.97

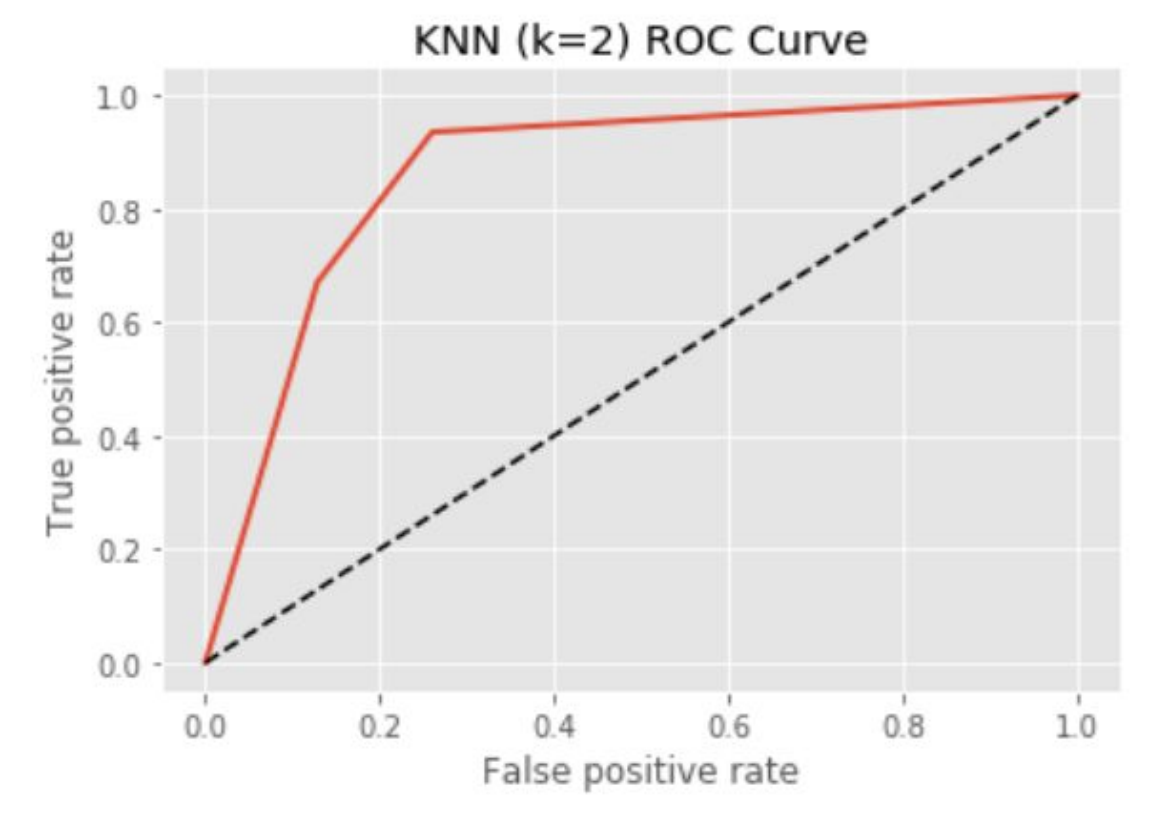
Logistic Regression



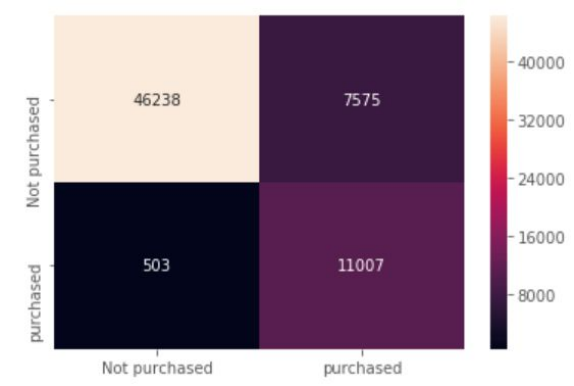


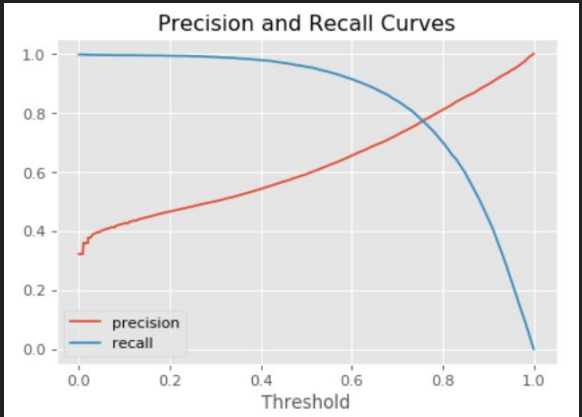
K-Nearest Neighbors (k=2)

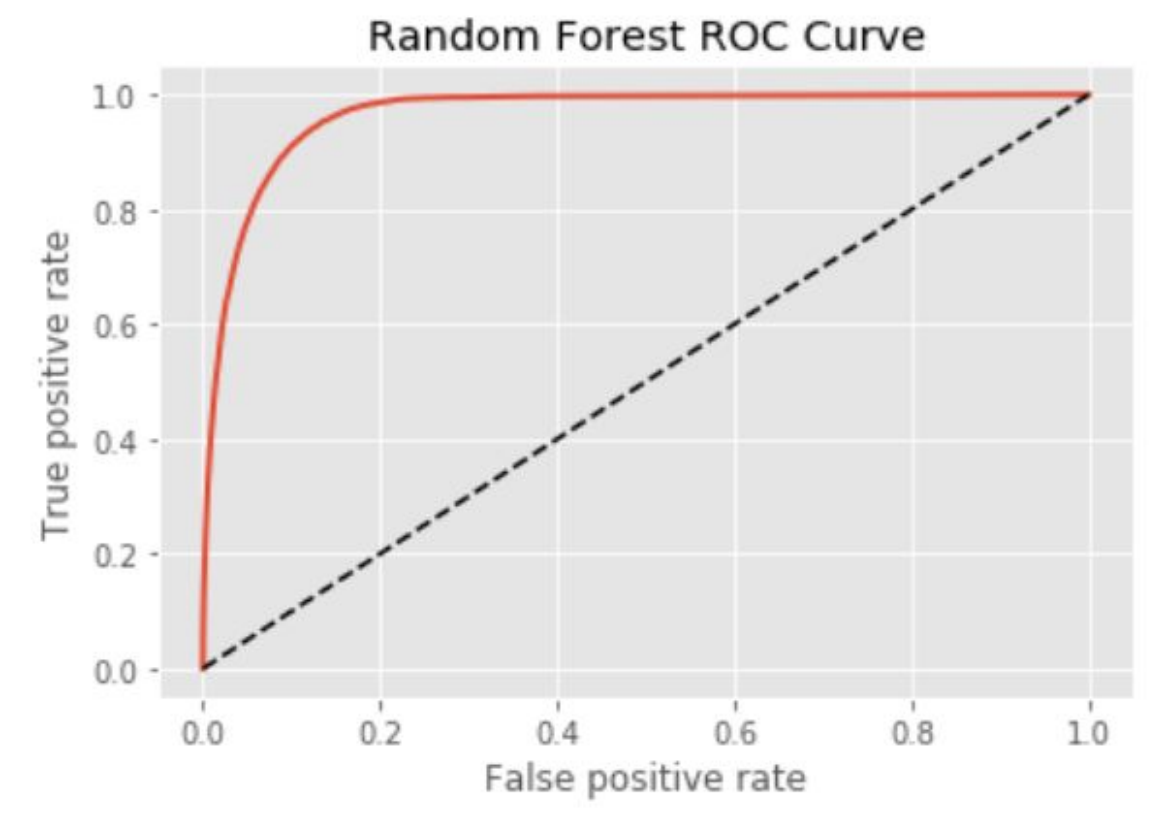




Random Forest

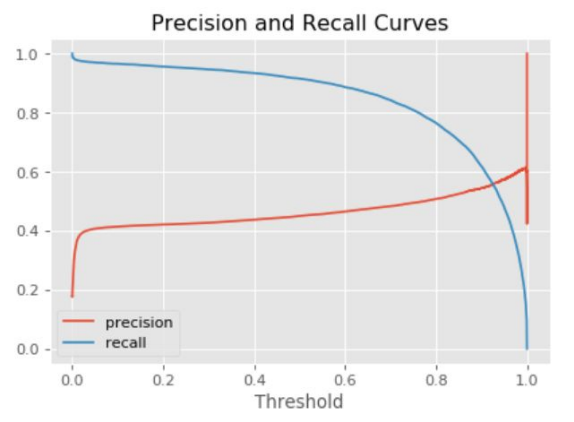


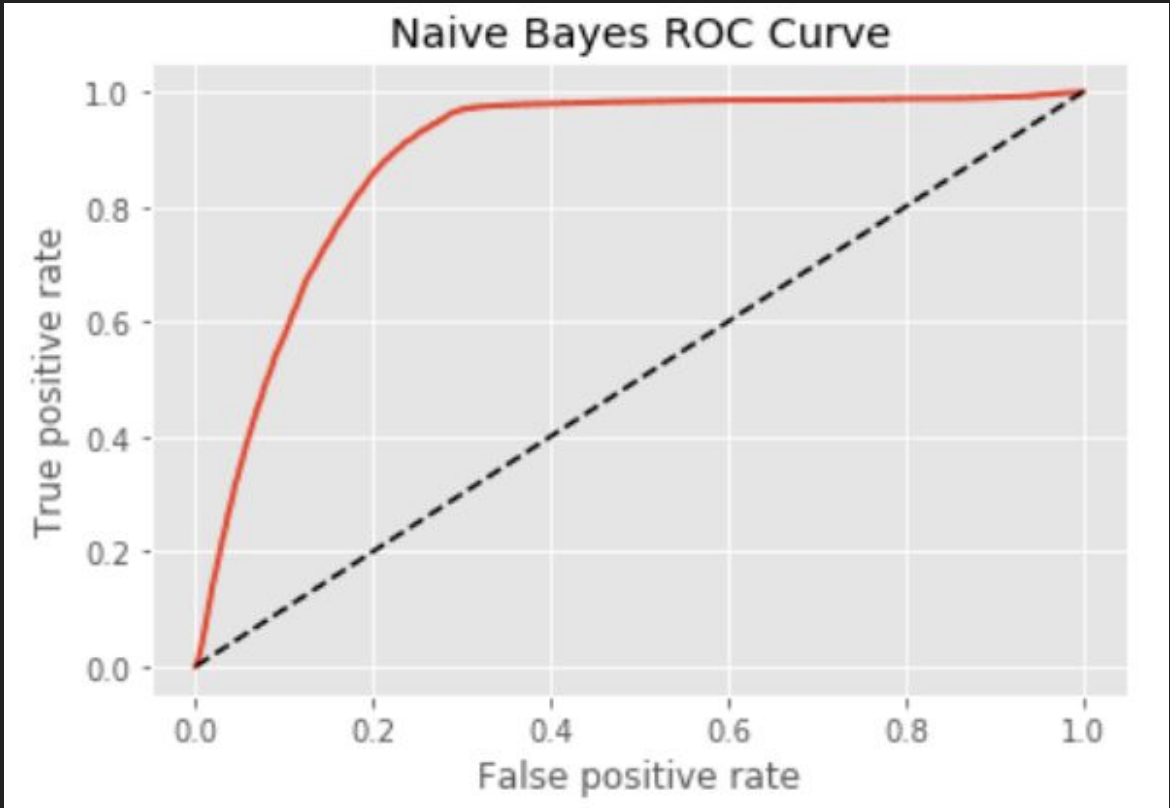




Naive Bayes

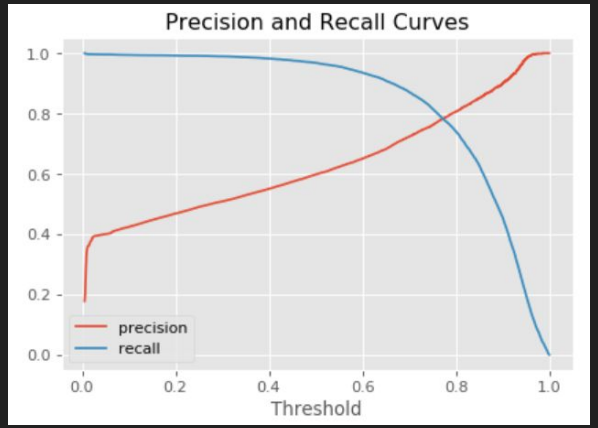

40935
12878
976
10534
Not purchased
purchased
Not purchased
purchased
40000
32000
24000
16000
8000

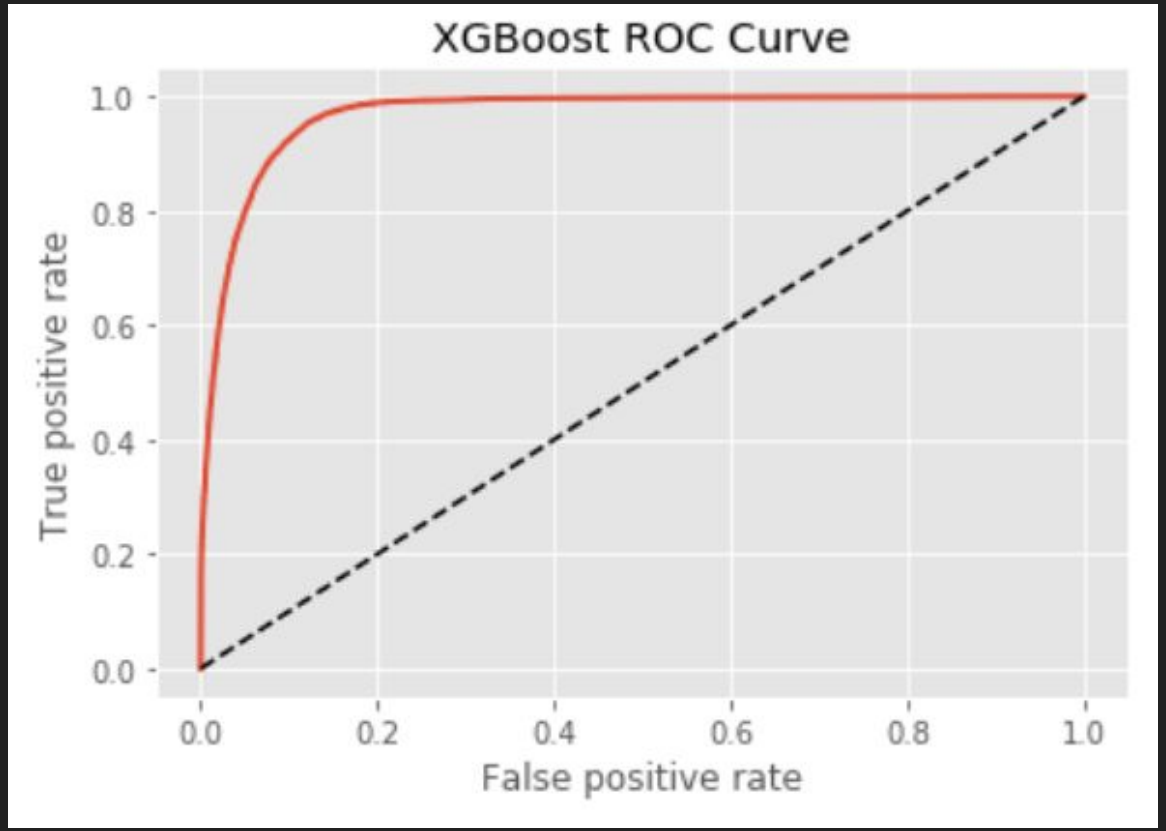

Precision and Recall Curves
precision
recall
Threshold
0.0
0.2
0.4
0.6
0.8
1.0


Naive Bayes ROC Curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

XGBoost


46321
7492
372
11138
Not purchased
purchased
Not purchased
purchased
40000
32000
24000
16000
8000


Precision and Recall Curves
precision
recall
Threshold
0.0
0.2
0.4
0.6
0.8
1.0


XGBoost ROC Curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0