Click or No Click?

CLICK THROUGH RATE ANALYSIS WITH MACHINE LEARNING

DS B19
LianYu Wang
Song Chang

Agenda:

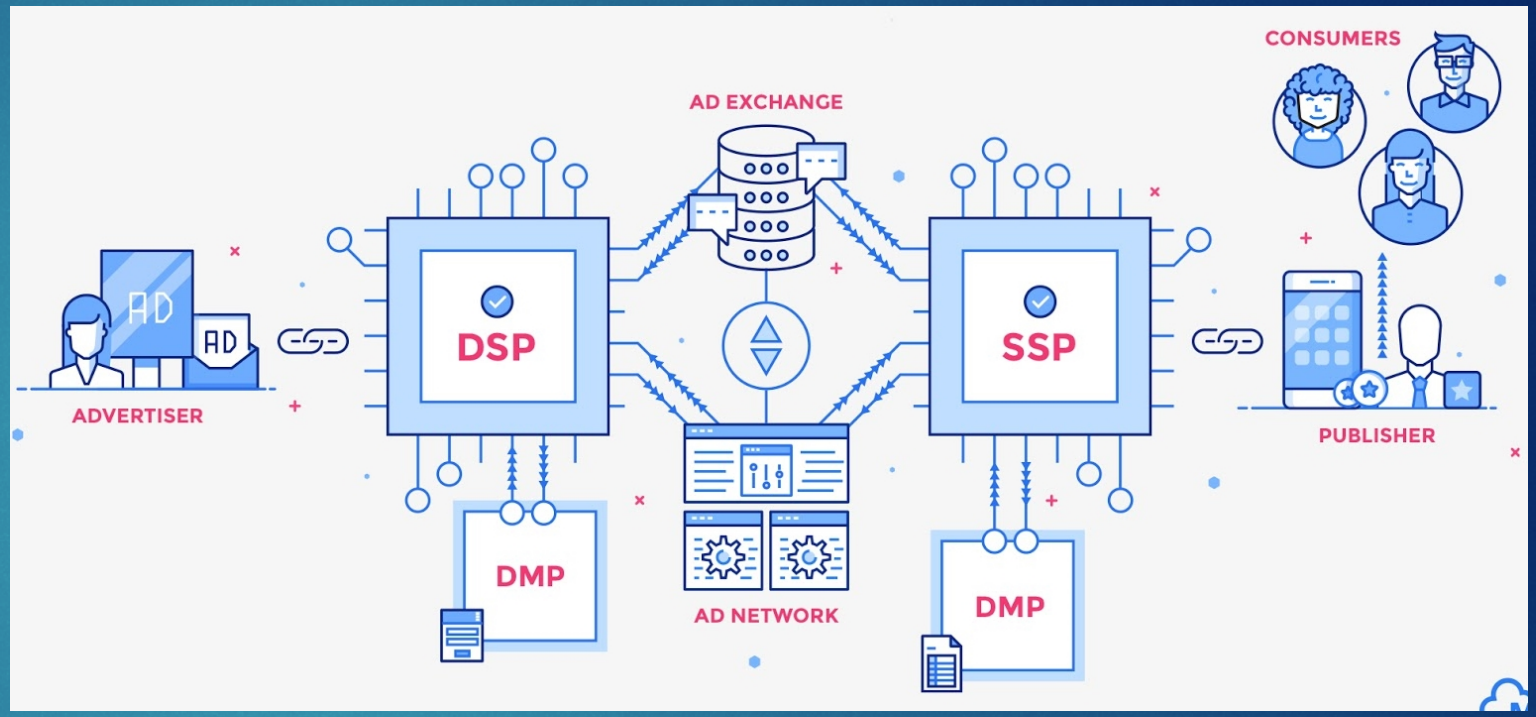
- Introduction to CTB
- Goal
- The Data
- Modelling
- Findings
- Challenges
- Conclusion
- Improvements

Introduction: Click or NoClick

- The world runs on clicks
- Attention is a product that is sold
- A generation of the worlds brightest minds have been manipulating your attention
- Clicks lead to micro-doses of dopamine
- Click Through Rate (CTR) is the metric that measures an Ad success rate
- Combined with web traffic and demgraphic data, a price can be determined
- Being able to accurately predict CTR will make your RTB (Real Time Bidder) more profitable whether you're sell / buyer

Goal: Accurately Predict Click

- To Do this to:
  - Understand the data
  - Clean & process the data
  - Build and test models
  - Analyze results
  - Make improvements
- Why is this important?
  - RTBs (Real Time Bidders)
  - Accuracy = profits



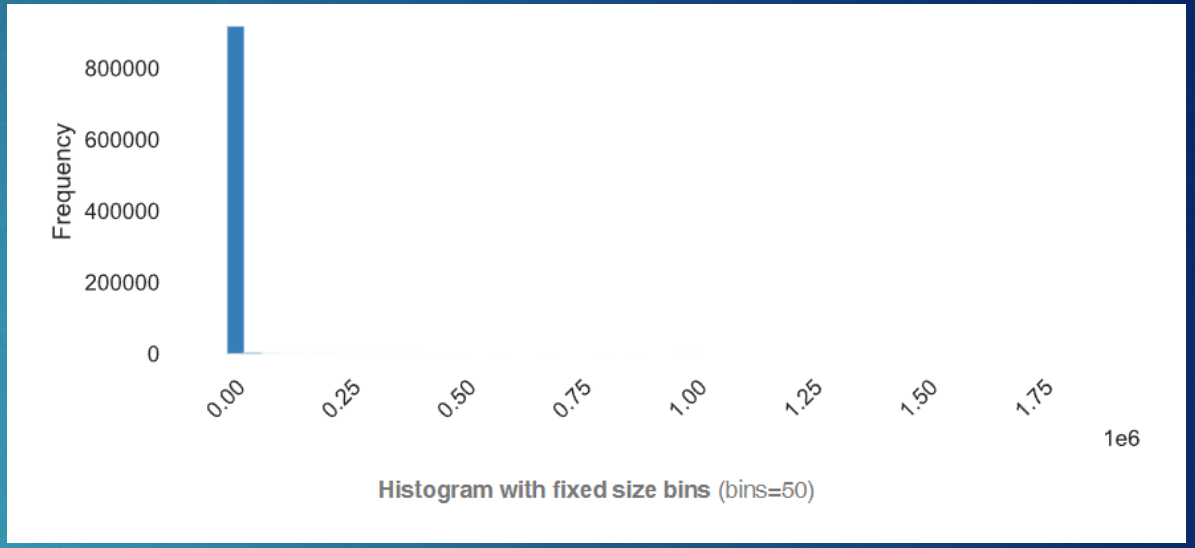
The Data: SQL / CSV -> Data Frame

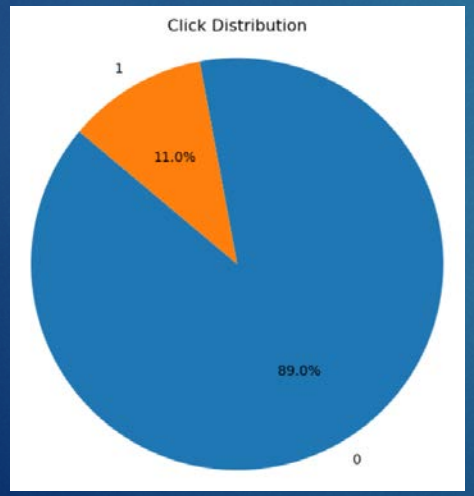
```
Zl2NO3,2017-08-04 12:35:01,1700,zrgM,CREDIT,jLX7,unknown,1000,male,深圳市龙华新区龙华街道清祥路1号宝能科技园乐康家居C1001A
pvpAZr,2017-08-04 12:35:05,2800,zO8g,CREDIT,PAv5R,unknown,1000,male,宝山区锦秋路135号
QOwLZ9,2017-08-04 12:35:02,100,zLGr,DEBIT,RW05,unknown,1000,female,天河区天河北路曜一城29号
w6Z6Wk,2017-08-04 12:35:05,1700,zO8g,CREDIT,DqVl,unknown,1000,male,福田区泰然七路210栋正宗桂林米粉
1N6o6D,2017-08-04 12:35:05,2600,4JBo,DEBIT,1GXN,3g+,1000,female,广州市天河区先烈东路252号
wWQeZM,2017-08-04 12:35:06,4200,4JBo,DEBIT,875E,3g+,1000,male,德胜内外大街11号院
kDkroK,2017-08-04 12:35:07,150,4JBo,DEBIT,Z7K5,4g,1225,male,北京市海淀区西平庄132号
0YVWYj,2017-08-04 12:35:02,800,zrgM,CREDIT,PJv5,unknown,1000,male,北京市朝阳区望京合生麒麟康馨美食城
QoABGX,2017-08-04 12:35:02,1900,zLGr,CREDIT,XPvq,unknown,1000,male,罗湖区田贝路2号化工大厦西首层10号
2JE730,2017-08-04 12:35:06,450,4JBo,DEBIT,1dpN,4g,1000,male,南京市师范大学中北学院食堂一层
7Gq2Y2,2017-08-04 12:35:05,1300,4JBo,DEBIT,kRMD,3g+,1000,male,深圳市福田区车公庙杭钢富春商务大厦广州餐厅
QOGG39,2017-08-04 12:35:07,690,zO8g,DEBIT,bppNK,unknown,1000,male,龙岗区平湖街道良白路白泥坑明辉集团一食堂
6Y5Z3W,2017-08-04 12:35:07,400,4JBo,DEBIT,dWOR,unknown,1,male,东城区北京火车站
```

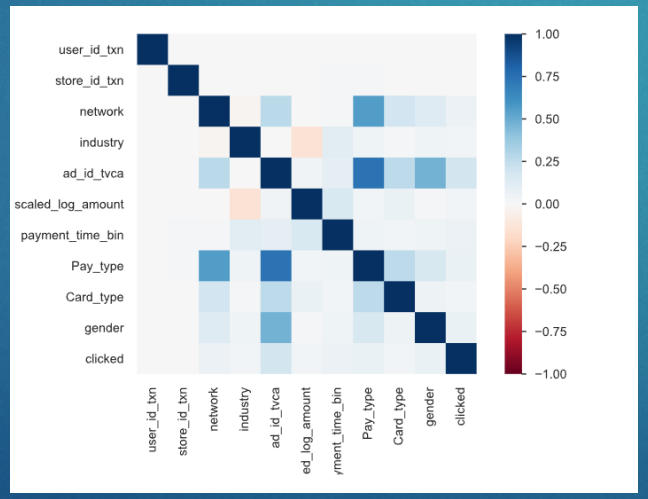
	user_id	payment_time	money	kind_pay	kind_card	store_id	network	industry	gender	view_time	click_time	ad_id	ad_loc	ad_label	clicked
1	000NK	2017-08-01 12:02:56	1600	4JBo	DEBIT	bpOLD	wifi	1000	female	2017-08-01 12:02:58	NaT	apjA	1.0	1001	0
2	001P2	2017-08-01 15:03:20	7810	4JBo	DEBIT	Kdkg6	4g	1225	male	2017-08-01 15:03:33	2017-08-01 15:05:26	apjA	1.0	1001	1
3	001RE	2017-08-01 11:54:37	1100	4JBo	DEBIT	VnOA	3g+	1000	female	2017-08-01 11:54:59	NaT	apjA	1.0	1001	0
4	001XE	2017-08-20 12:21:42	3000	zLGr	CREDIT	pkPk	unknown	1000	male	2017-08-20 12:21:45	NaT	zj9k	2.0	1009	0
6	002mK	2017-08-20 13:34:33	1800	4JBo	DEBIT	67jj1	3g	1000	female	2017-08-20 13:34:43	NaT	a9PI	1.0	1009	0
9	005KA	2017-08-01 18:35:12	6400	4JBo	DEBIT	LprL8	3g+	1000	female	2017-08-01 18:36:25	2017-08-01 18:36:28	apjA	1.0	1001	1
10	005KA	2017-08-20 18:31:50	2100	4JBo	DEBIT	KN2O	3g+	1000	female	2017-08-20 18:32:05	NaT	a9PI	1.0	1009	0

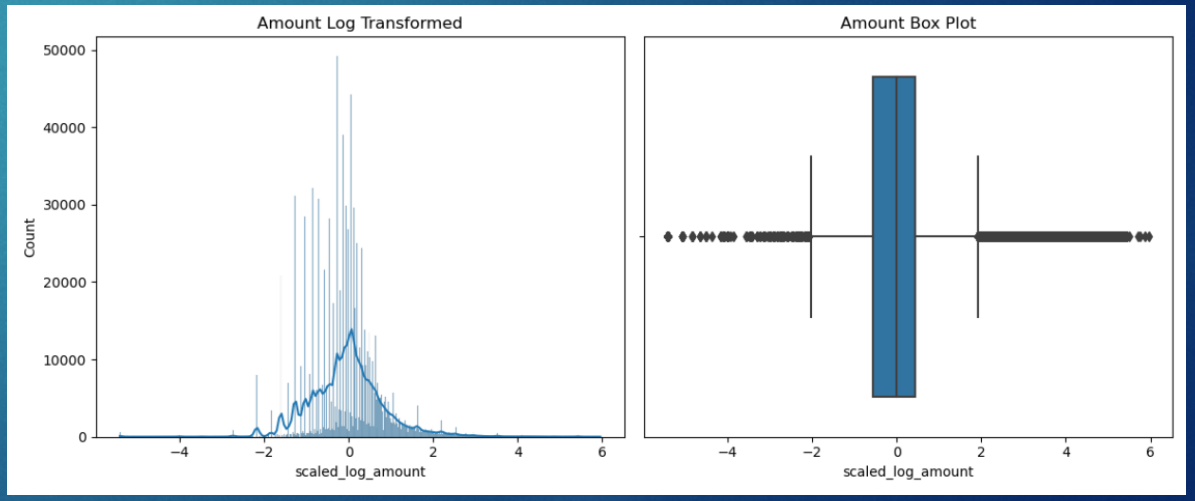
The Data: Some Notes

- Clicks = 0 majority Class
- One day of data
- Duplicates (6380 0.6%)
- Skews and outliers (‘money’)
- Data is Imbalance (ROS, RUS, SMOTE)
- Feature Engineering (‘Time of Day’)







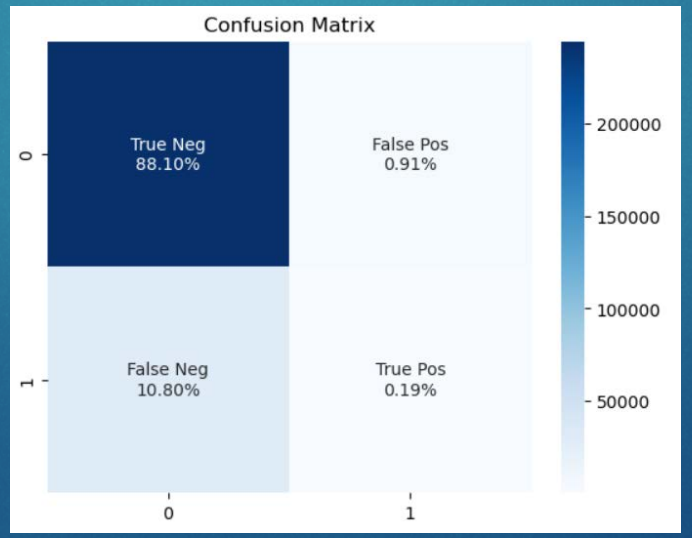


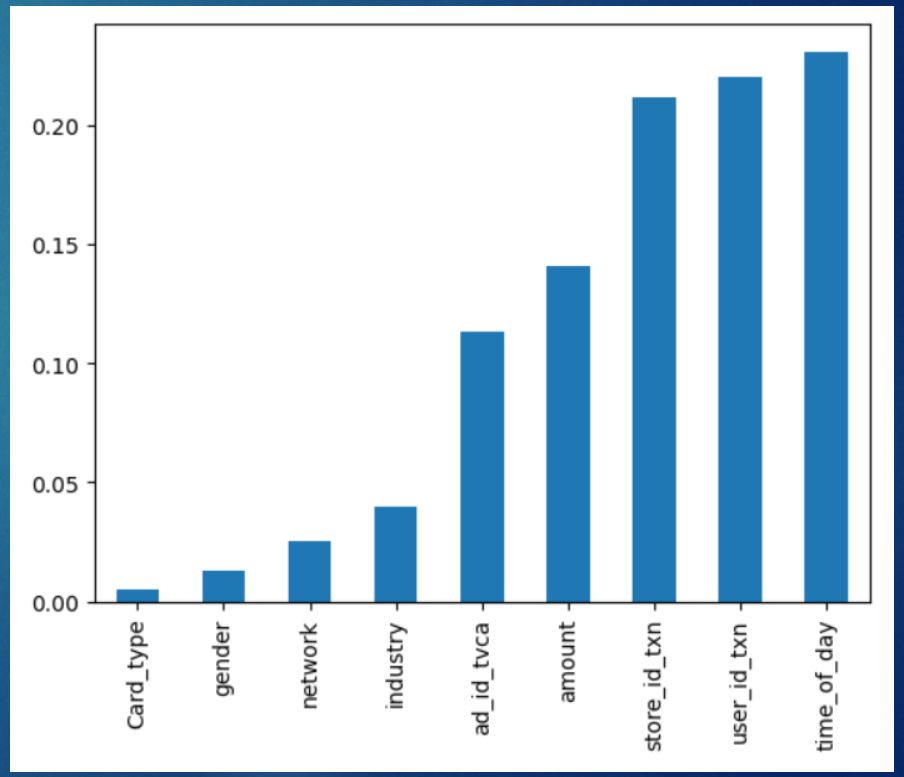
Findings: DTC -> Random Forest

- Logistic Regression (LR) performed very poorly after balancing test data set
- DTC did better than LR
- Random Forest did even better
- Random Oversampling > SMOTE > RUS
- Did not finish running hyper-tuning

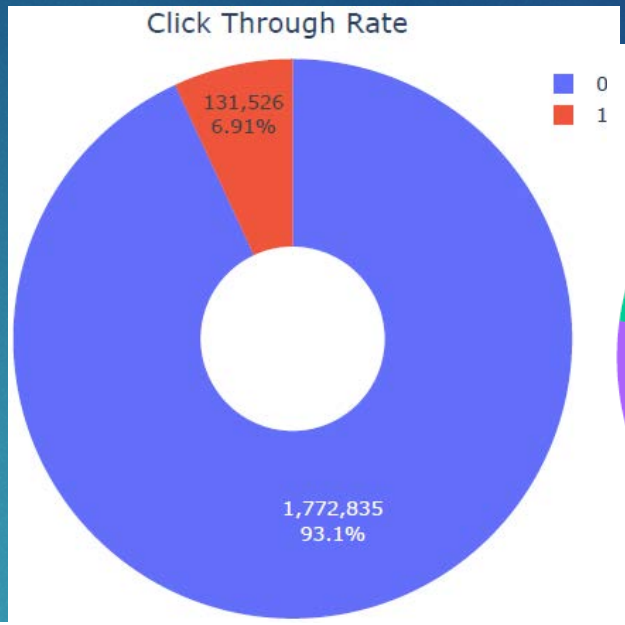
```
Accuracy: 0.8829021149234625
Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.99      0.94    246839
           1       0.17      0.02      0.03     30476
```

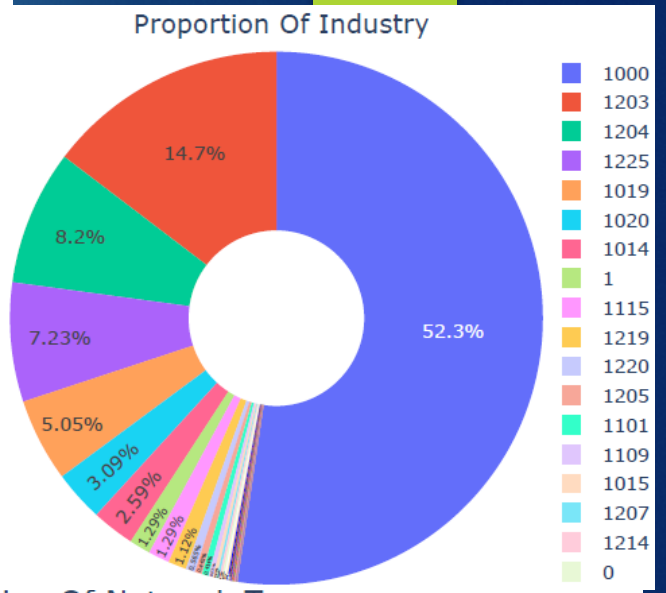
	fit_time	score_time	test_f1	test_accuracy
0	65.467301	1.561091	0.002928	0.889494
1	67.036812	1.582780	0.002265	0.888791
2	66.494469	1.563190	0.002260	0.888556
3	65.502227	1.549207	0.003202	0.887727
4	65.315944	1.558864	0.002681	0.892685

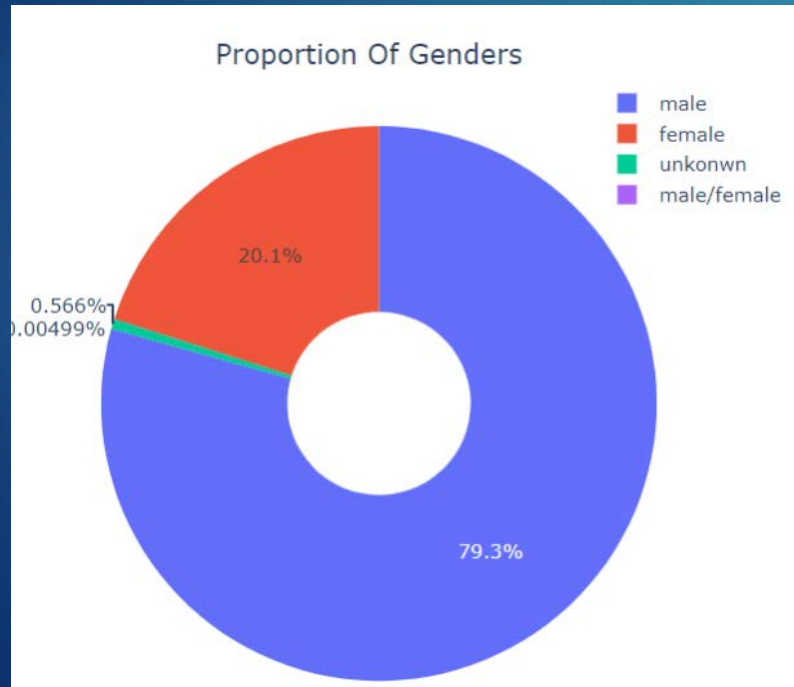


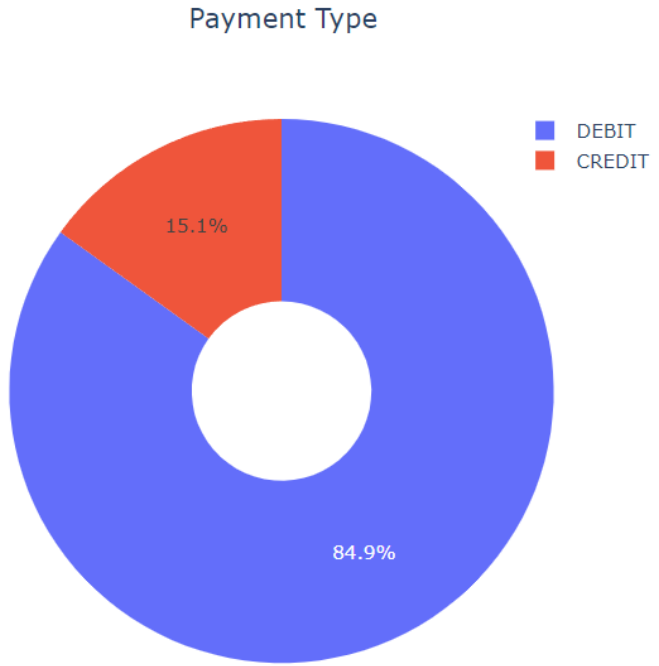


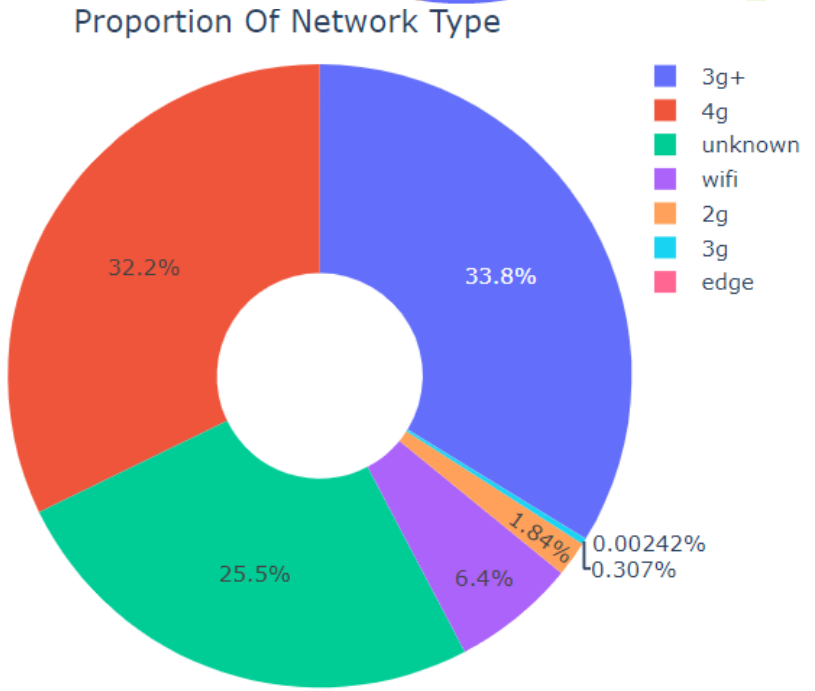
The Data: SQL (4day)



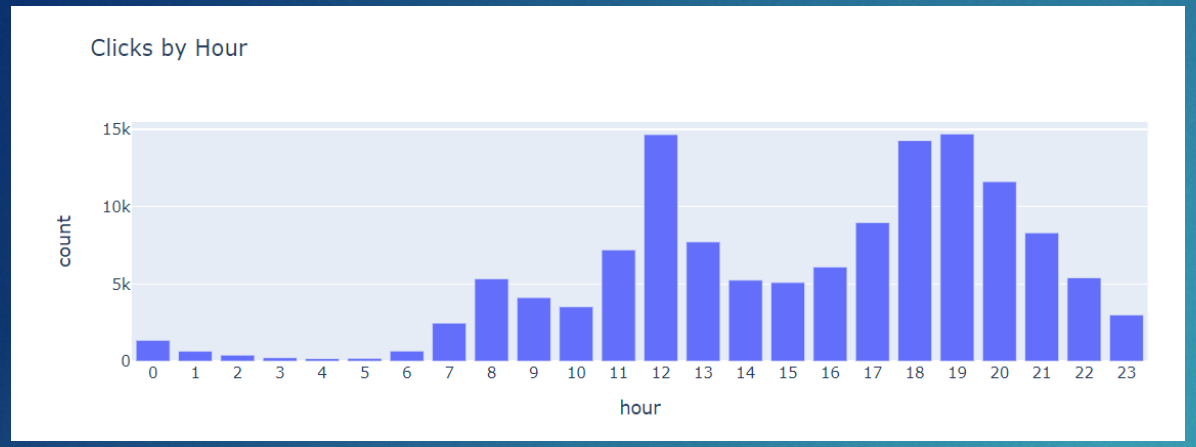


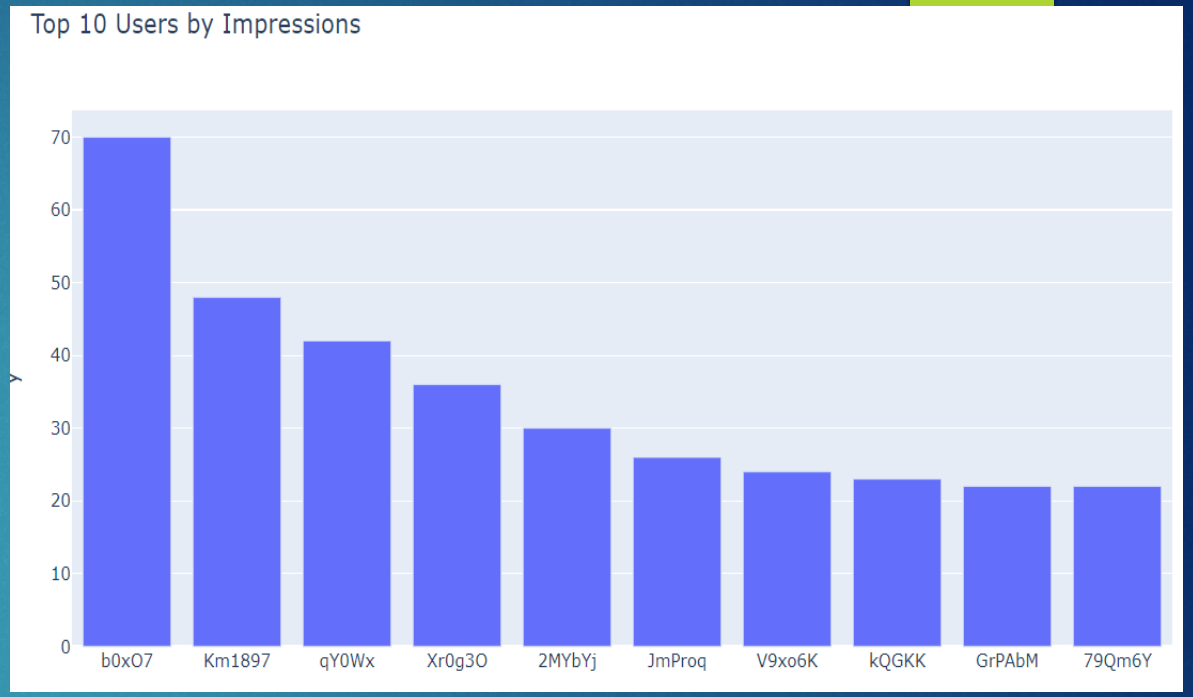


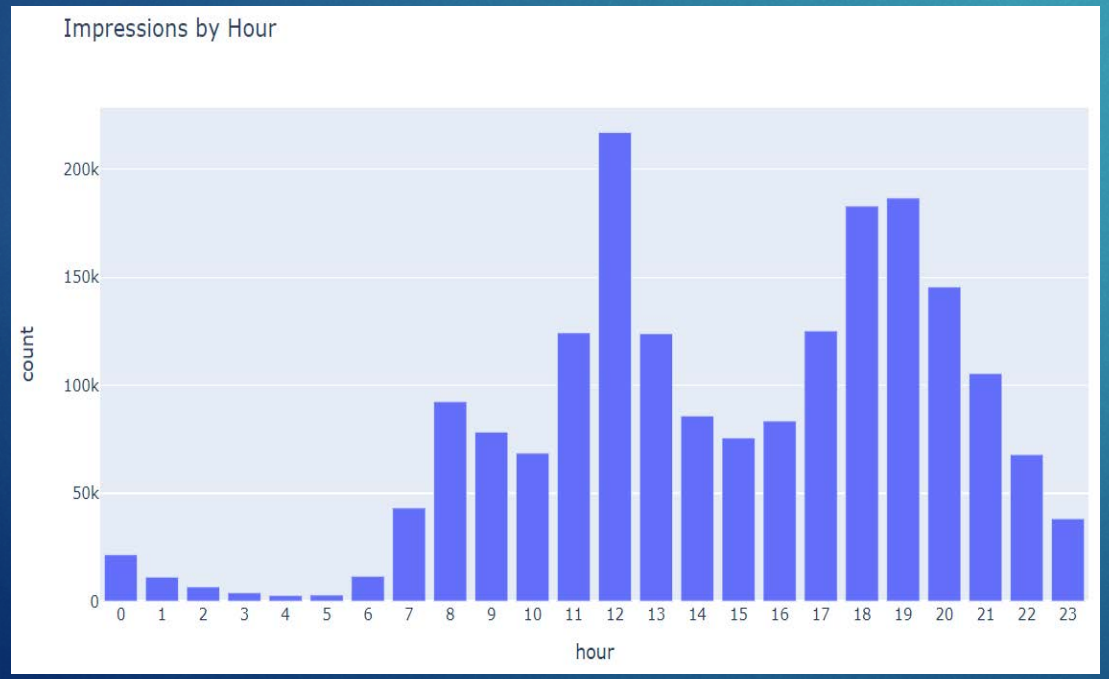


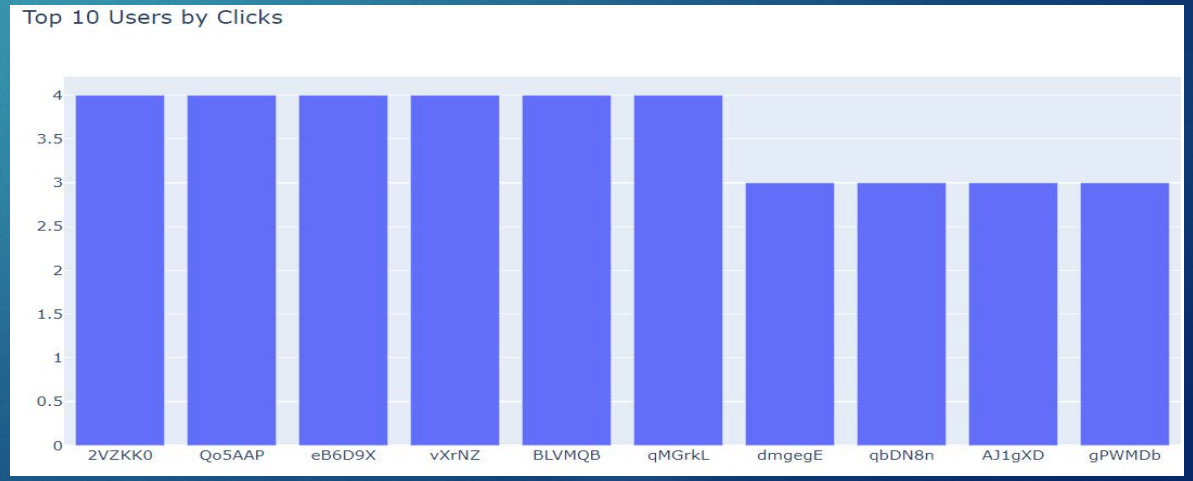


The Data: SQL (4day)









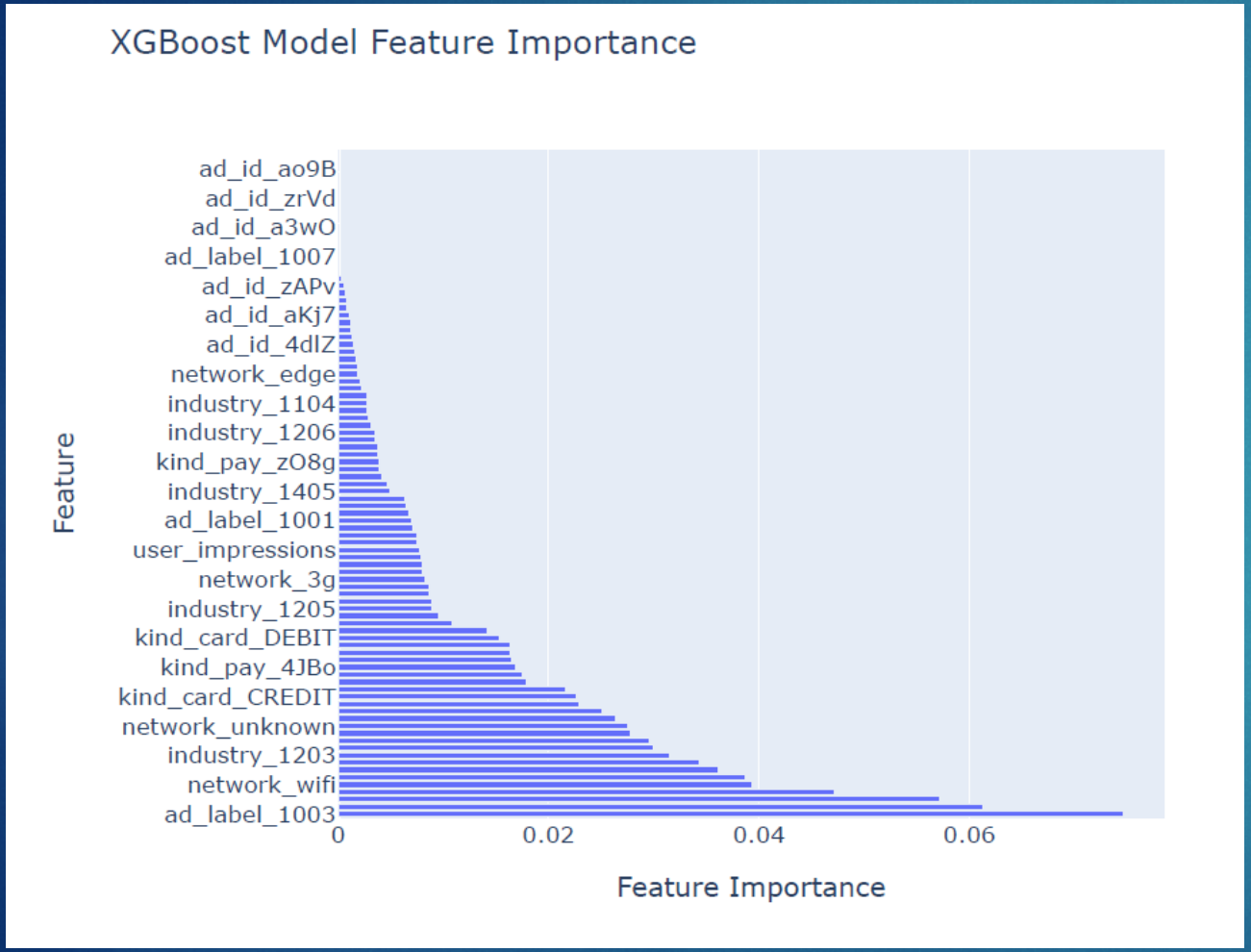
Model: XG Boost

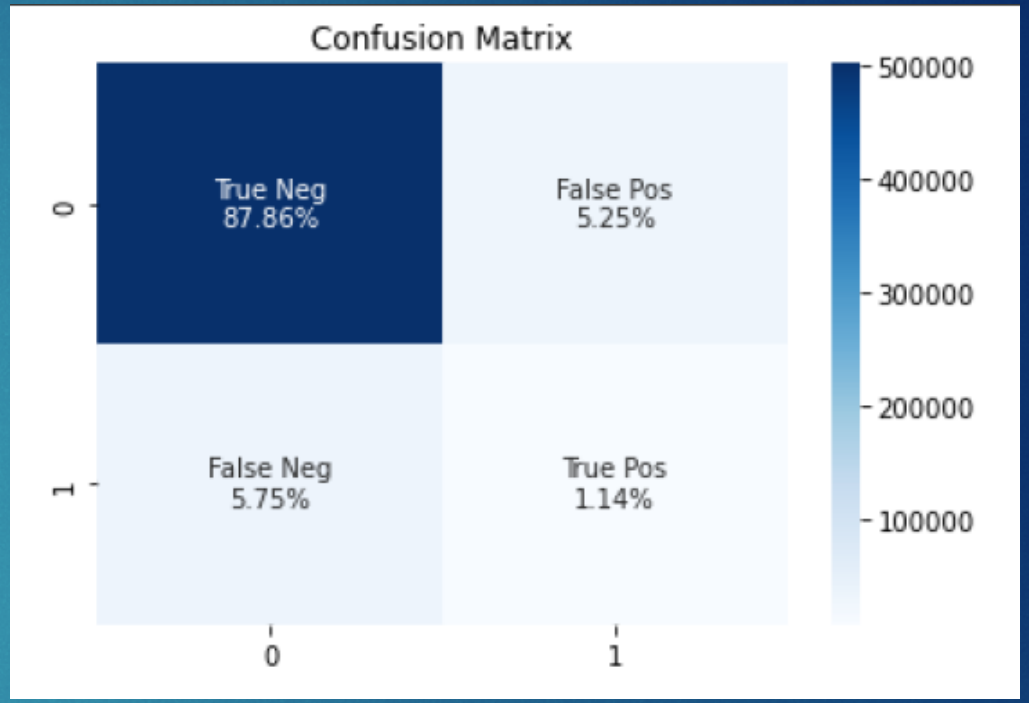
- 5-Fold cross validation

	fit_time	score_time	test_f1	test_accuracy
0	5.652780	2.914221	0.000760	0.930992
1	5.463806	0.189017	0.001265	0.930887
2	5.006375	0.190014	0.000000	0.931351
3	5.148387	0.194015	0.000000	0.931263
4	5.105383	0.182016	0.001014	0.931000

```
accuracy: 0.89
precision: 0.18
recall: 0.17
```

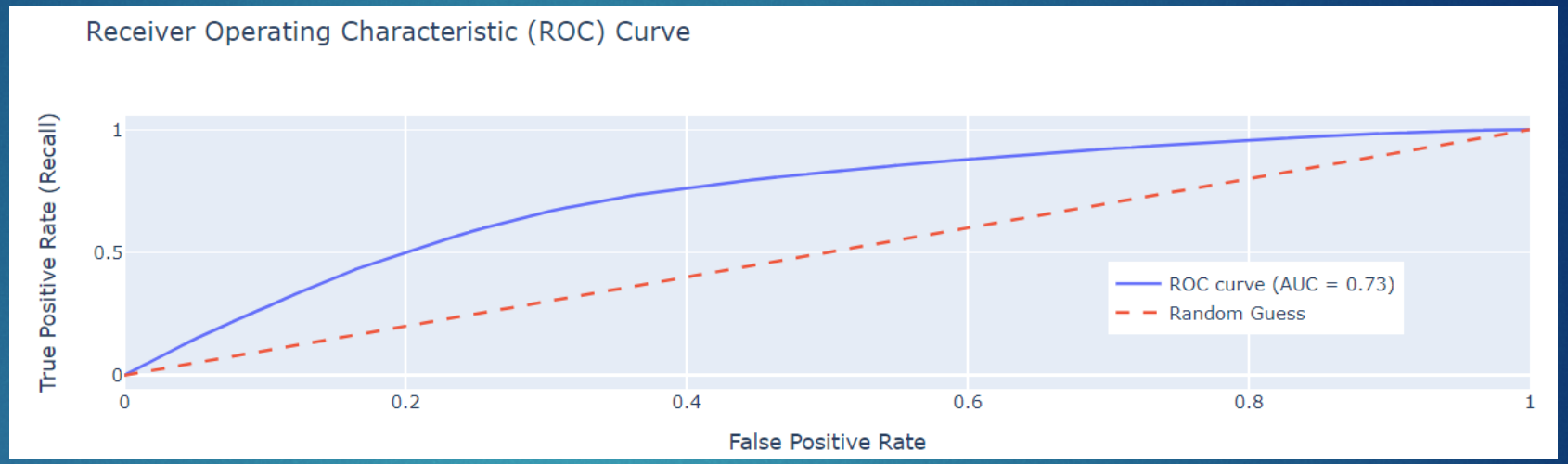
Model: XG Boost



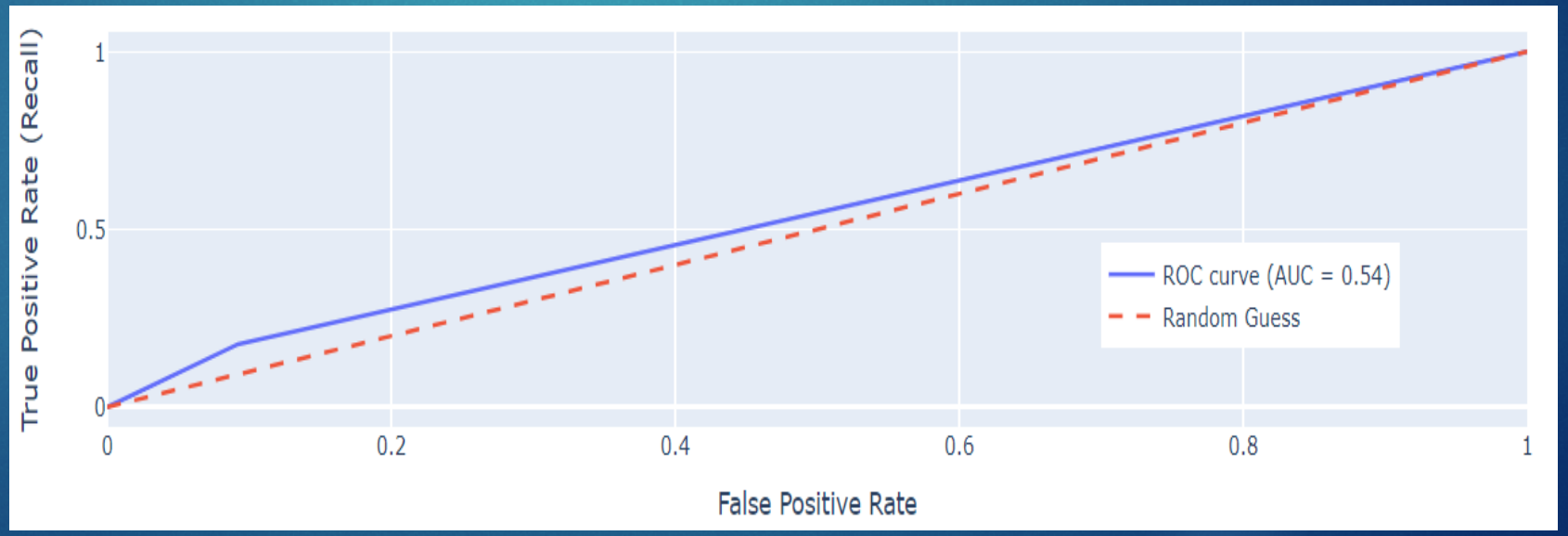


Findings: ROC Curve

XGBoost



DecisionTree



Challenges:

- Large dataset requires powerful computer
- 1 or 4 day dataset cannot capture patterns of greater time scale (seasonal, monthly, pay day)
- Hyperparameter tuning takes a lot of time
- Many concepts to process and apply
- Lack of experience

Conclusion :

- Random Forest and Xgboost both give good results. But model performance was ultimately constrained by the quality of data.
- In conclusion is our model has high accuracy, but low precision and recall, high true negative, but very very low true positive, So our model can correctly predict someone will not click the ad after viewing it, but can't accurately predict someone will click the ad after viewing it. I think this is because of the dataset is highly imbalanced. Even oversampling cannot overcome the negative bias caused by a lack of successful clicked through samples.
  So to improve our model, we need to collect more samples of people who clicked the ad after viewing it.

Improvements : More *.*

- More Samples
- More parameter and feature testing
- More practice with projects & exercises
- More Compute
- More RAM
- More Time

Thank You!

LIANGYU@LIANGYU.COM