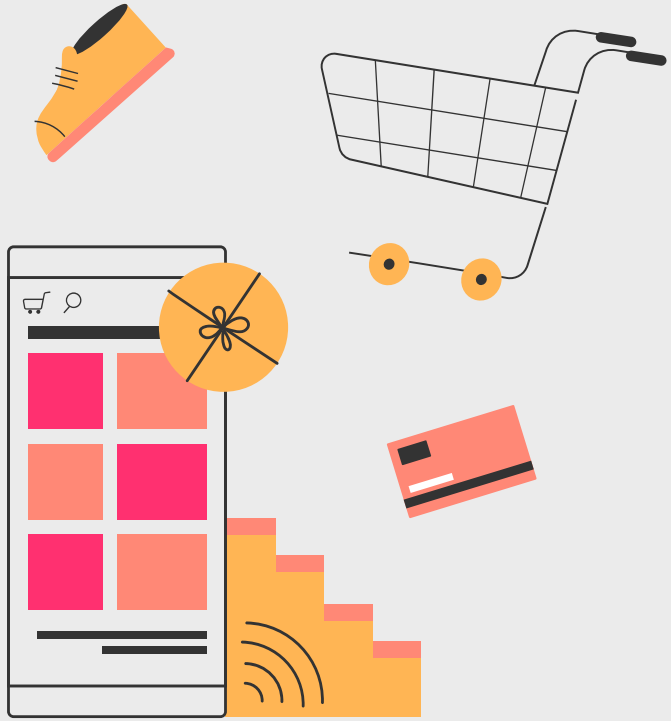
Real-time customer purchasing tendency prediction

SharkyData

Buyan Li, Yun Lin, Yu-Yuan Chang, Yafei Dong

fingerhut®

Table of contents



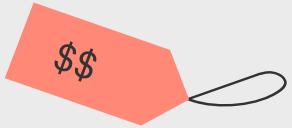
01

Introduction

Our Focus

Predicting the likelihood that customer will purchase during the shopping section in real-time

Analyzing the customer behaviors: page views, search history, and promo-code usage ect.



The implications for businesses seeking to optimize their e-commerce strategy

02

Exploration & Visualizations

Yun Lin

Initial Selection

From the original dataset, we selected 17 features:

'visitid','pageurl','pagename', 'pageeventvar2',

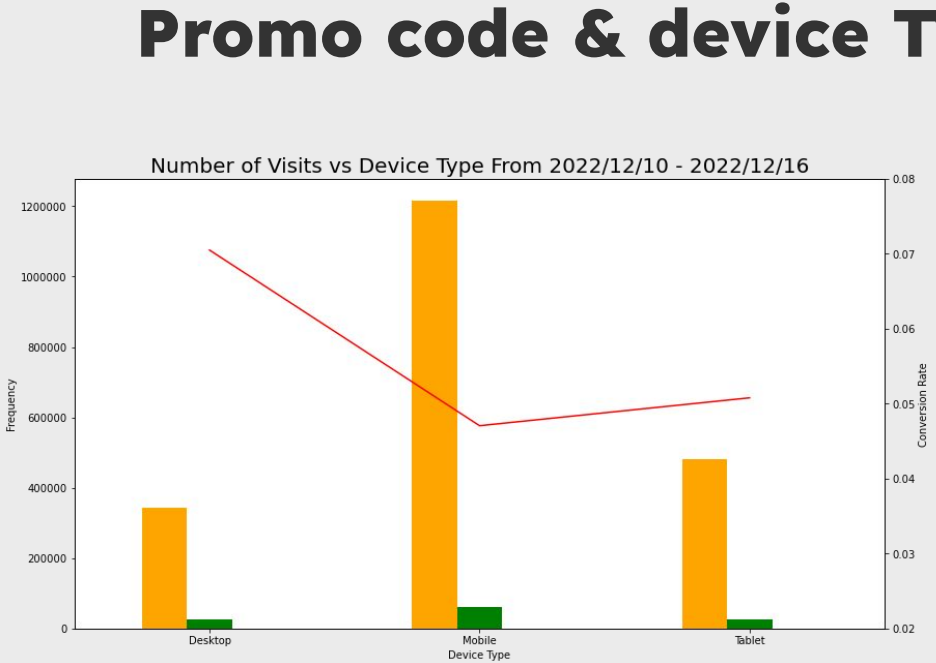
'pagetype', 'visitdatetime', 'hit_time_gmt',

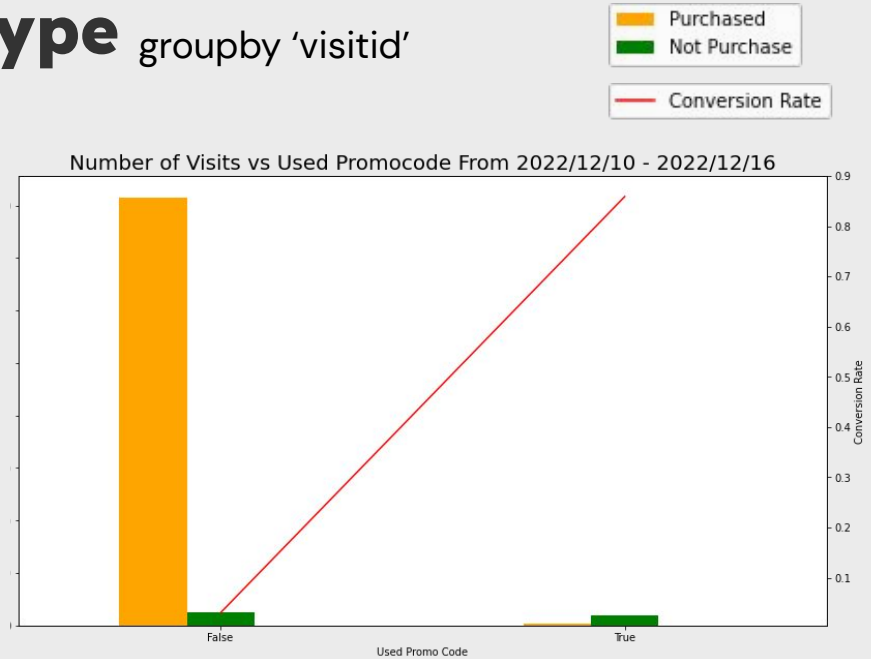
'productlist', 'searchterms', 'searchresults',

'newvisit', 'post_evar27', 'evar28', 'evar83',

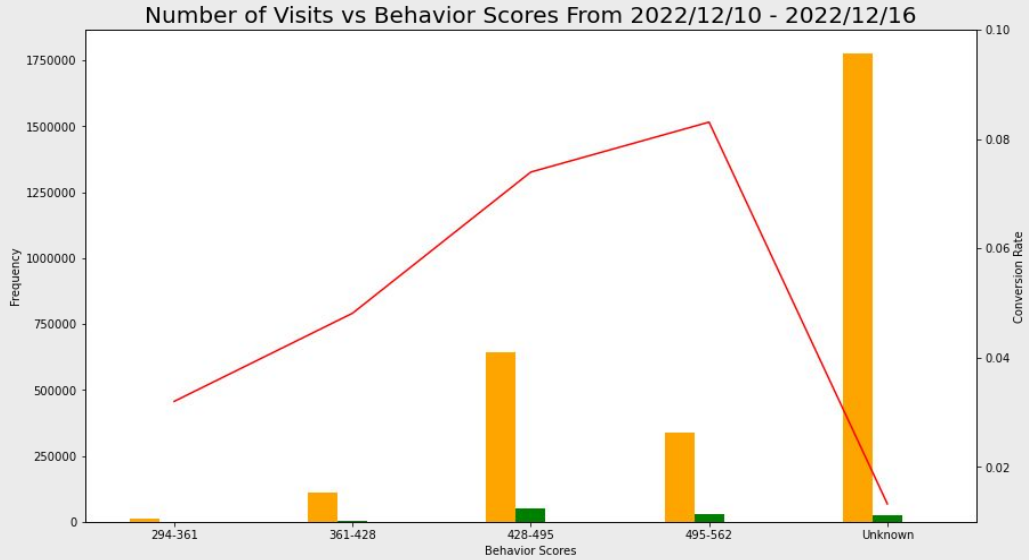
'promocode', 'devicetype', and 'ordernumber'

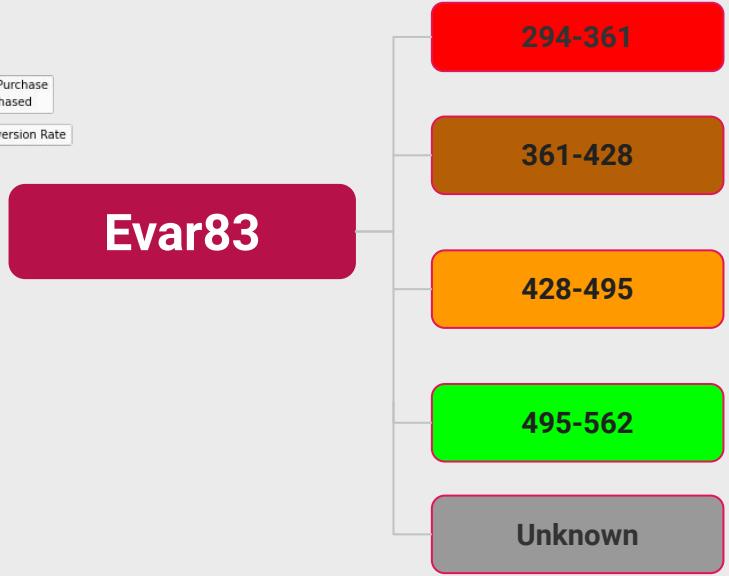
Promo code & device Type groupby ‘visitid’





Behavior Score (Evar83)





03

Feature Engineering

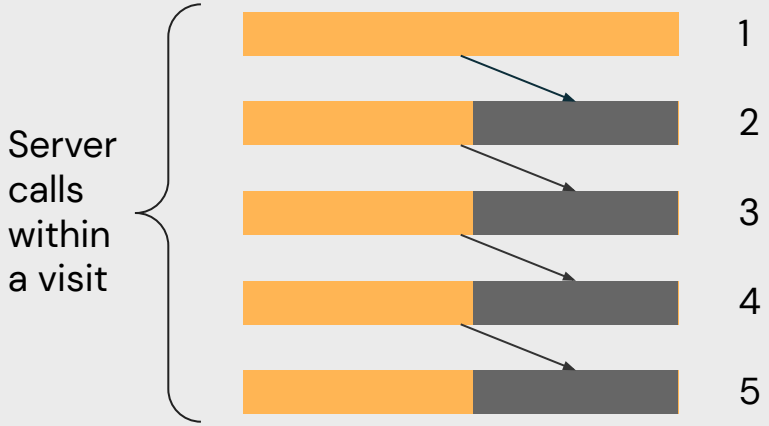
Adding memory features

- What are “memory” Features and why do we need them?
    - Store information from previous server calls within the same visit
    - Needed for generating a prediction for each server call

Adding memory features

Before adding memory features

After adding memory features



Adding memory features

Memory features added:

`sum_cart`	`sum_prod`
`prod_in_cart`	`prod_viewed`
`avg_cart`	`avg_prod`
`min_cart`	`min_prod`
`max_cart`	`max_prod`

Adding other features

Other features added:

`Timedelta:` Time between this server call and the previous one

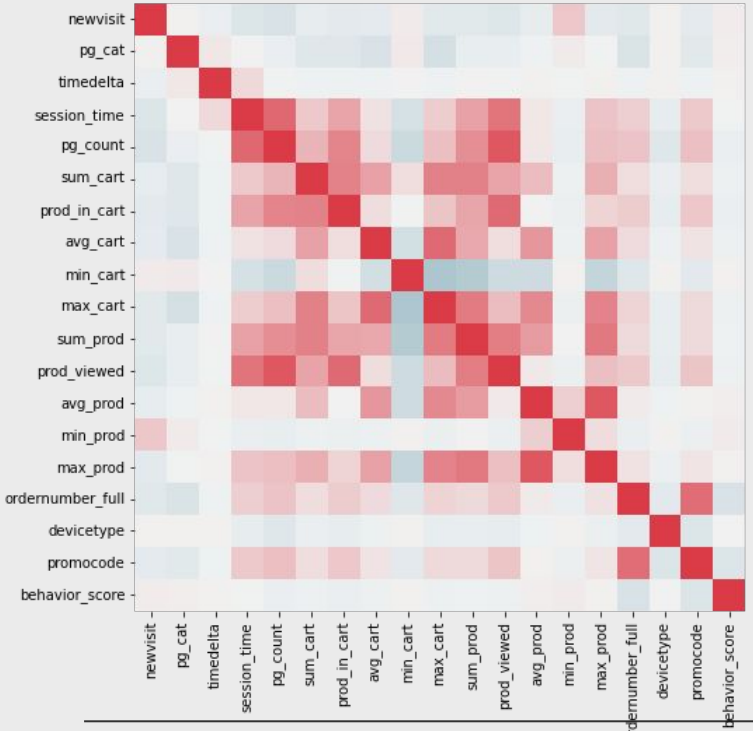
`Pg_cat:` Category of the current page

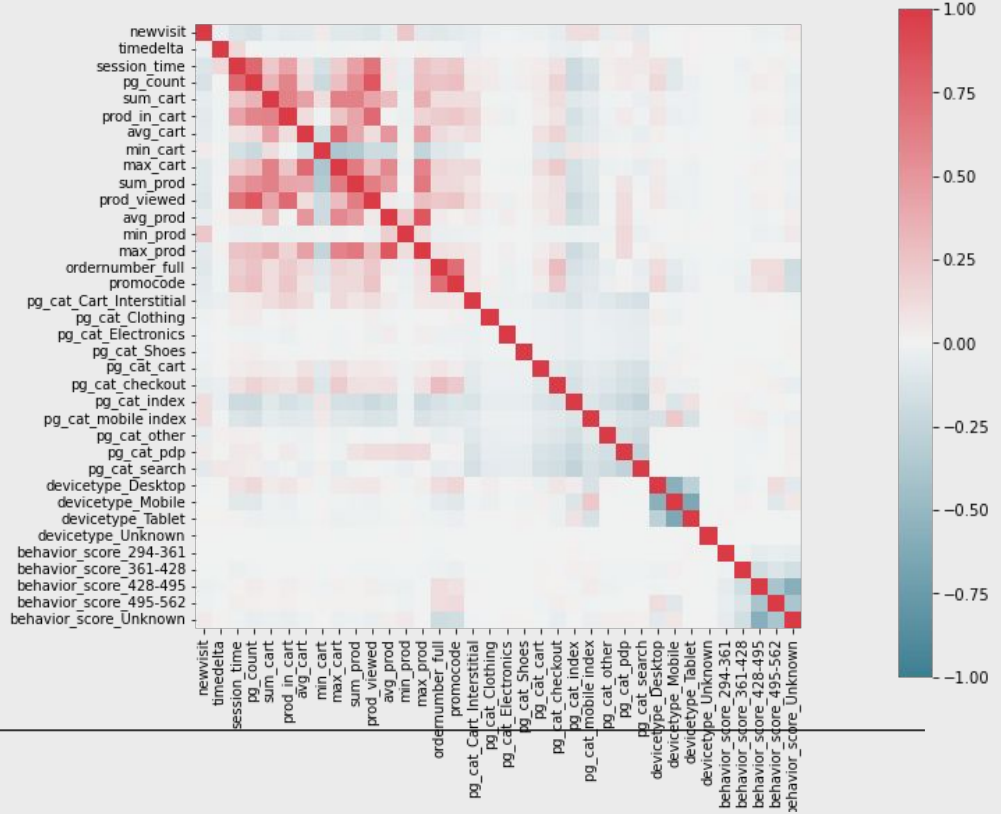
`Session_time:` Total time spent since the start of this visit

`Pg_count:` Total server calls generated since the start of this visit

Correlation matrix

before & after one-hot encoding





04

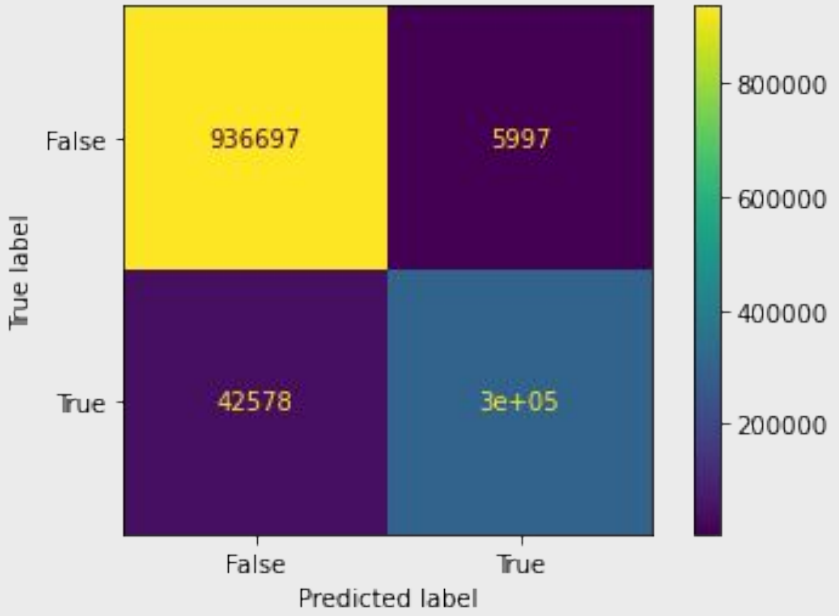
Model Building & Evaluation

Model Building

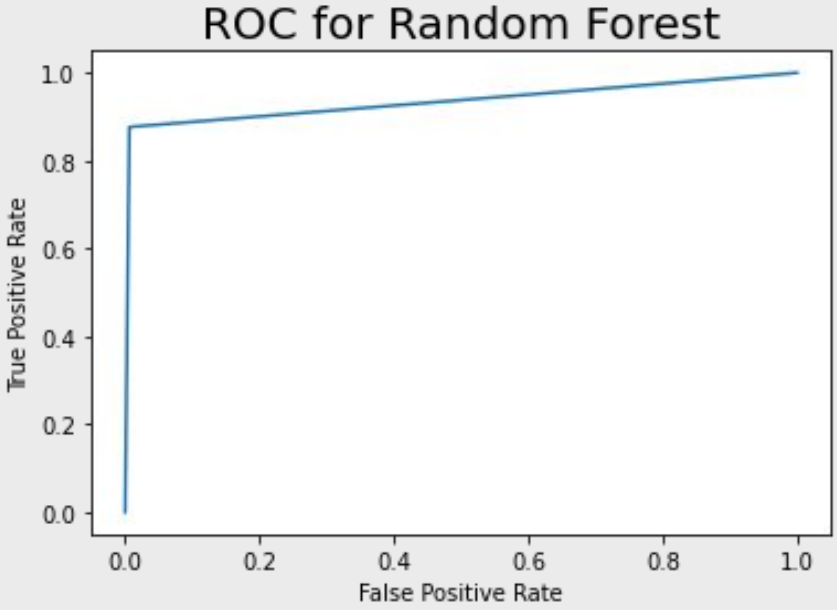
1. Random forest
2. Gradient Boosting Classifier
3. Logistic Regression
4. SVM
5. KNN

Random Forest

Accuracy:	0.9624
Precision:	0.9806
Recall:	0.8765
F1-Score:	0.9256



Random Forest

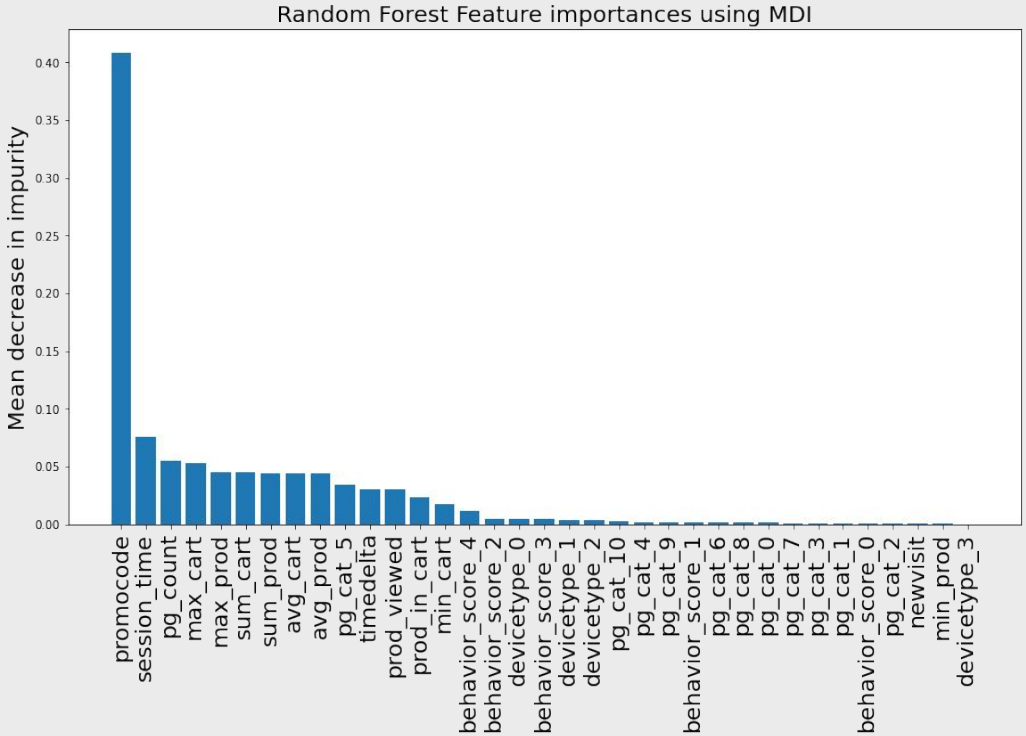


1. **Parameter tuning**:
   criterion: “entropy” and “gini”.
   max_depth: [10, 20, 30]

2. **Cross-Validation:**
   5 folds

3. Performance the best!!

Result: Feature Importances

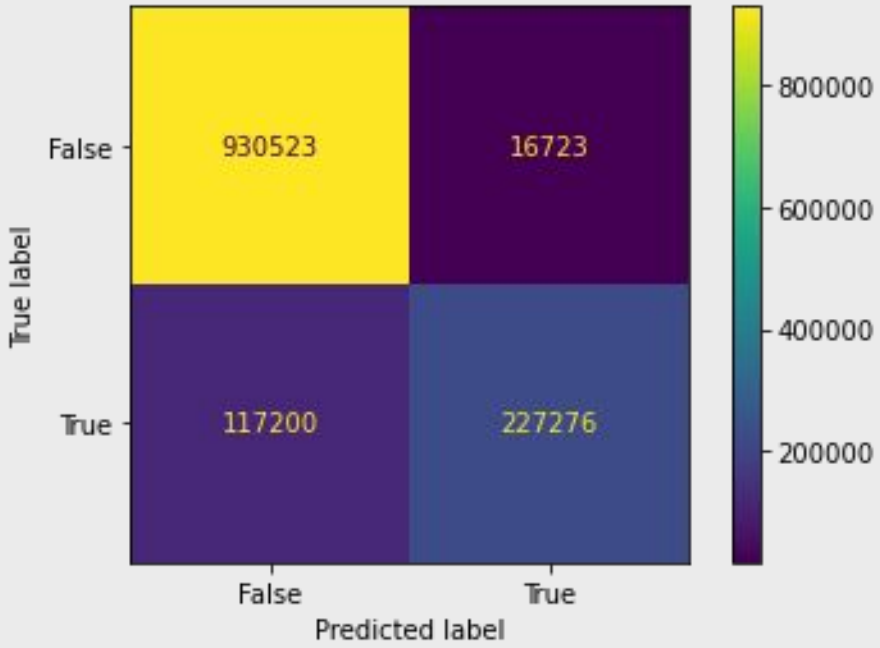


Feature Importance:

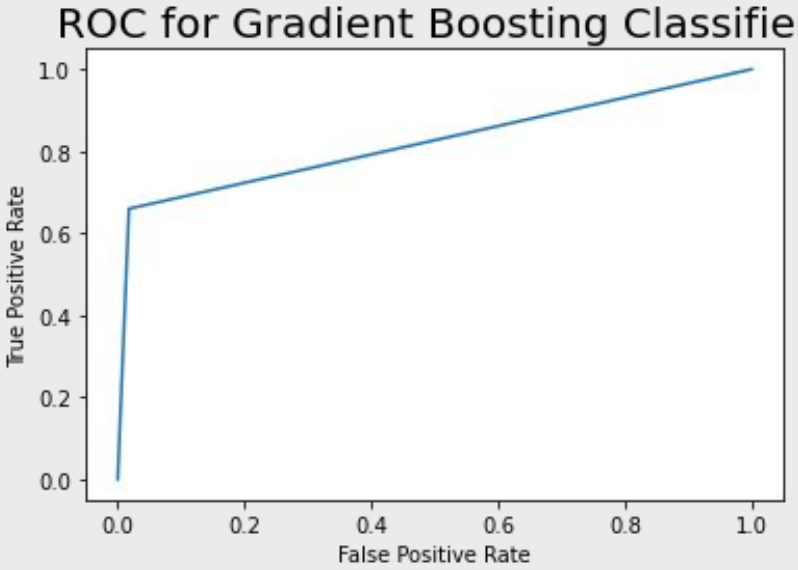
1. **Promo code**
2. **Session time**
3. **Page count (Number of action)**

Gradient Boosting Classifier

Accuracy	0.8963
Precision	0.9314
Recall	0.6597
F1-Score	0.7724



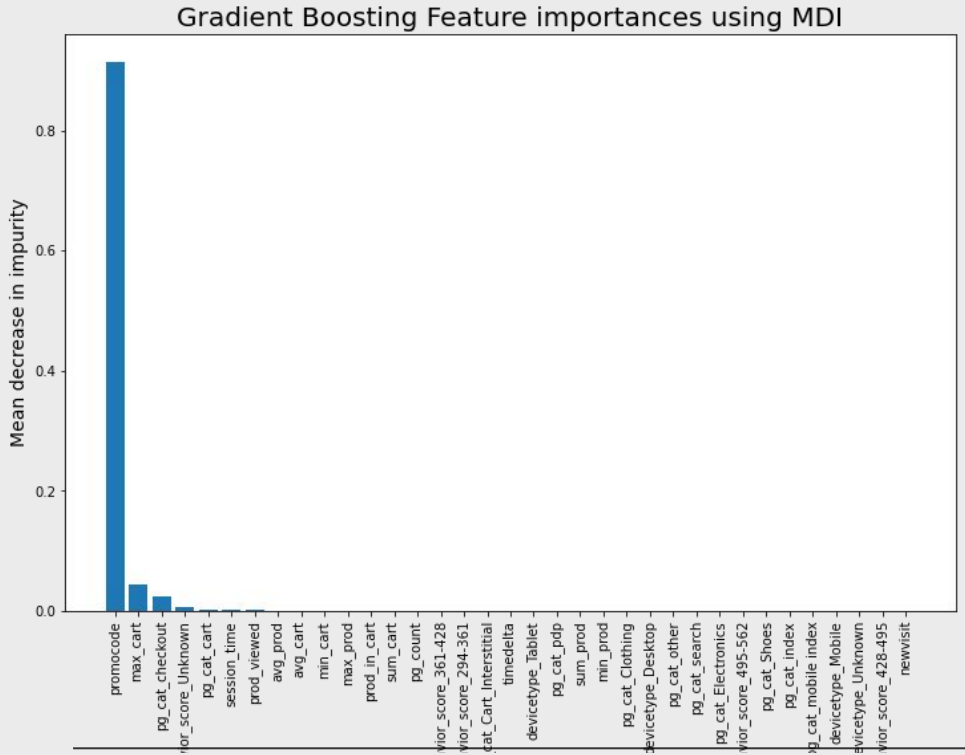
Gradient Boosting Classifier



1. **Parameter tuning**:
   Learning rate: [0.01, 0.1, 0.5, 1]
   n_estimator: [10, 50, 100, 150, 200]

2. **Cross-Validation:**
   5 folds

GB: Feature Importances

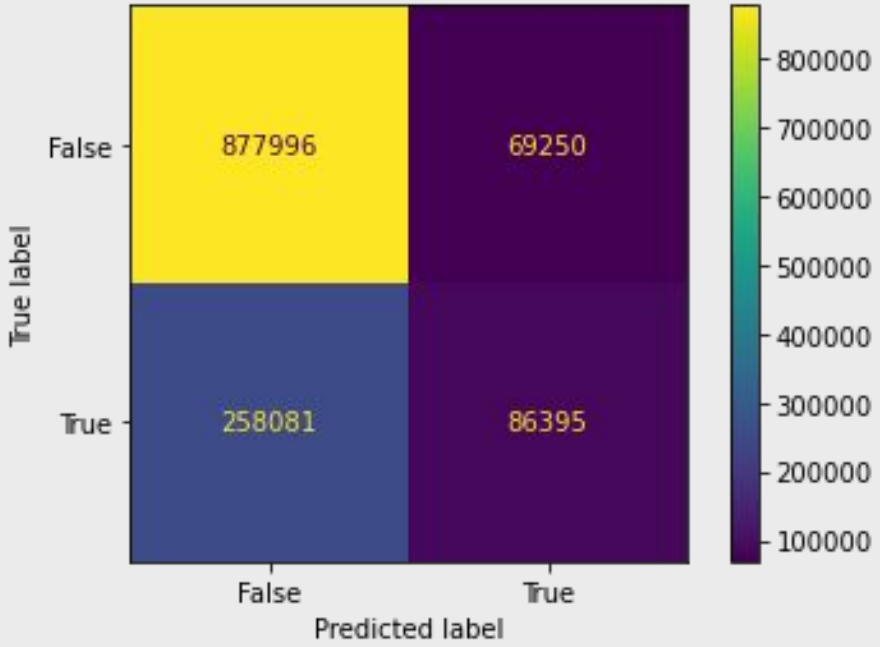


Feature Importance:

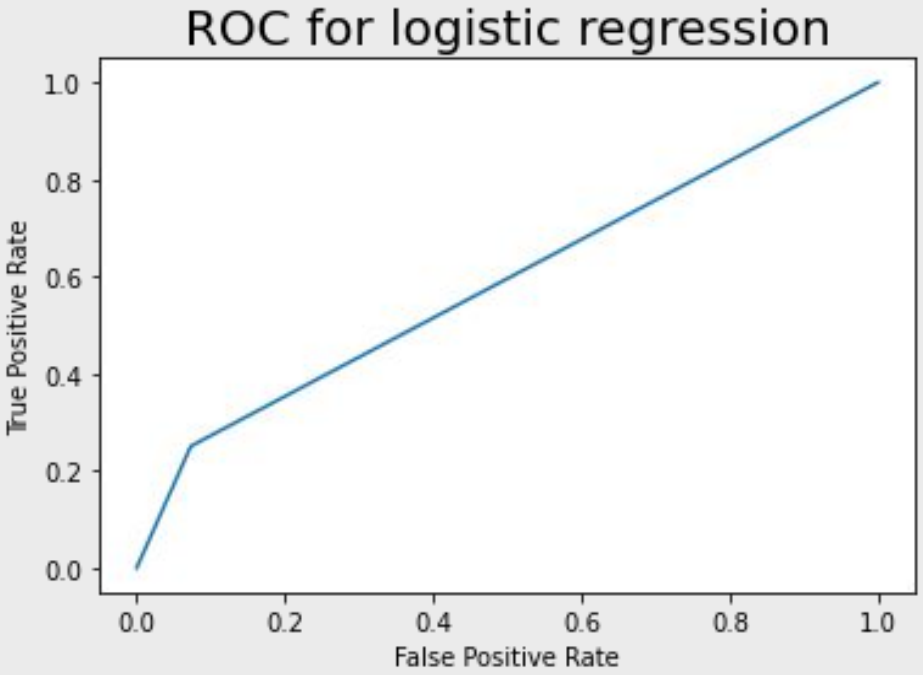
1. **Promo code**
2. **max_cart**
3. **pg_cat_checkout (Number of action)**

Logistic Regression

Accuracy	0.7465
Precision	0.5550
Recall	0.2508
F1-Score	0.3454



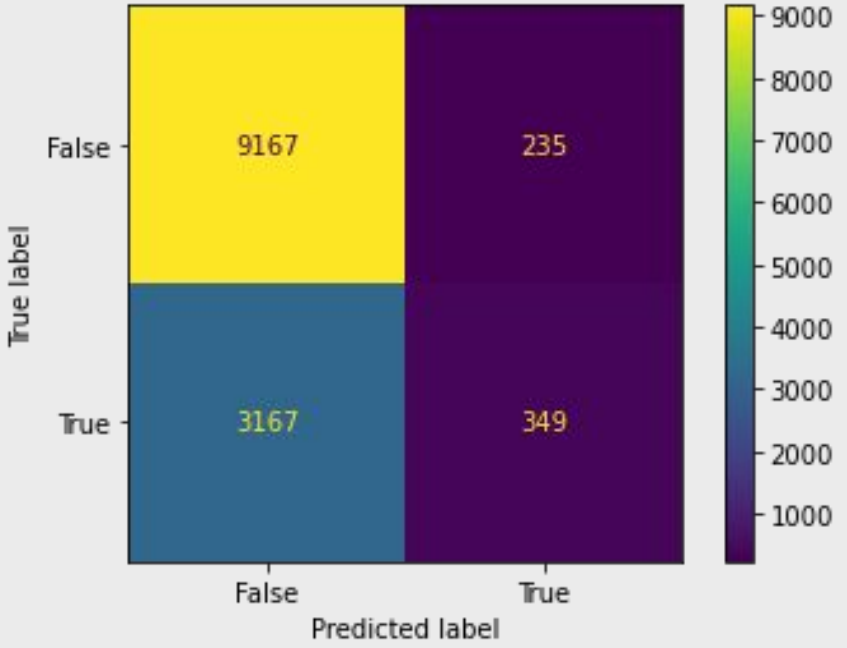
Logistic Regression



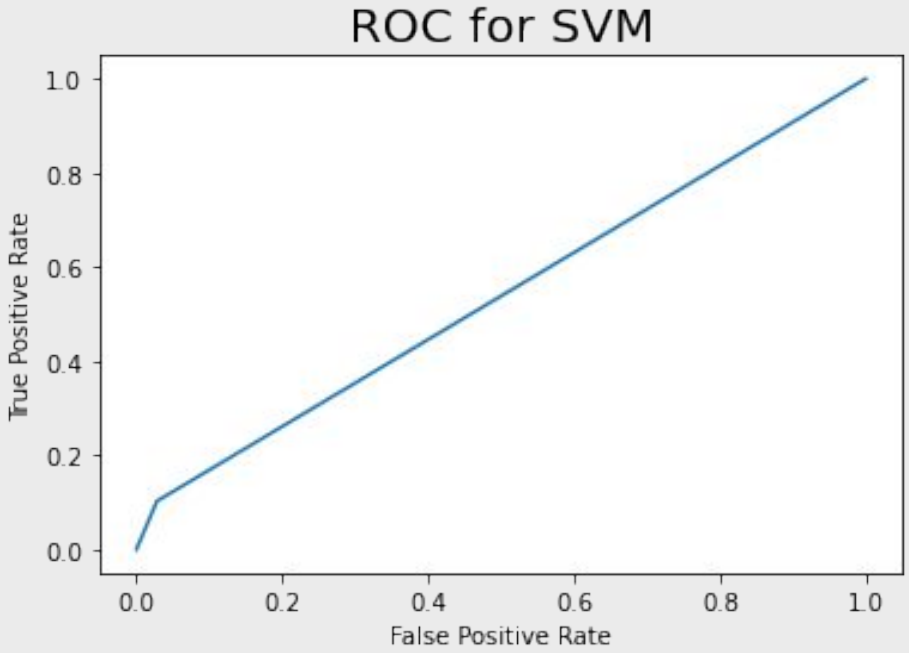
1. **Parameter tuning**:
   penality: ‘l1’, ‘l2’, ‘elasticnet’, None
2. **Cross-Validation:**
   5 folds
3. Did not performance well

SVM

Accuracy	0.7366
Precision	0.5976
Recall	0.0992
F1-Score	0.1702



SVM



1. **Parameter tuning**:
   kernel: “linear”, “poly” and “rbf”
2. **Cross-Validation:**
   5 folds
3. Only use 1% of the sample.
4. Did not performance well

Summary

1. Analyzing the frequency of pageviews for a particular product.
2. Design new feature and construct model.
3. Being able to provide insights into customer preferences and forecast which products are likely to be purchased in the future.

Future Work

1. Factors external: economic conditions, weather, promo code and social.
2. Recommendation system
3. We can limited our model into first few click/pages

Limitation

1. Features Mapping were not consistent
2. Not sure when ‘Promocode’ applied
3. The data is during Holiday season
4. The data contains only one week timeframe

Thanks for listening

Backup slides

NewVisit→ NewVisitor

	visitid	newvisit
0	001	0
2	001	0
3	001	0
4	001	0
5	001	0
...	...	...
6458604	99983884244834782820100728265624489661 8	1
6458605	99997363407775132790292379433452812751	1
6458606	99997363407775132790292379433452812751	0
6458607	99997363407775132790292379433452812751	0

	visitid	newvisit
0	001	0
2	001	0
3	001	0
4	001	0
5	001	0
...	...	...
6458604	99983884244834782820100728265624489661 8	1
6458605	99997363407775132790292379433452812751	1
6458606	99997363407775132790292379433452812751	1
6458607	99997363407775132790292379433452812751	1

Promocode → checkout?

So we assume promo code applied in cart

s.eVar82 | Promo Code Errors in Checkout (eVar 82)

Errors submitting a promo code in checkout

eVar82 is empty

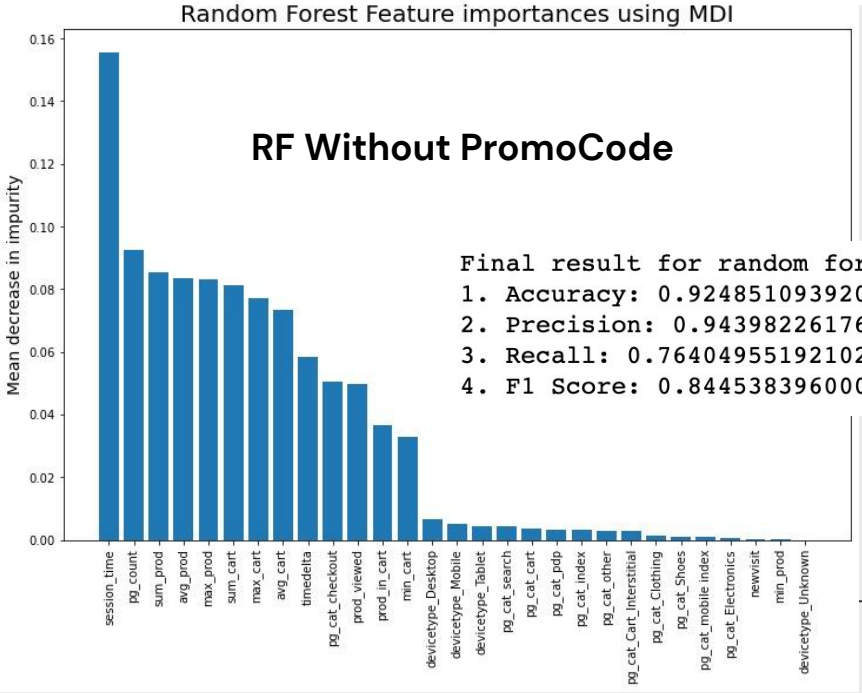
2041286 rows × 4 columns

```
[184]: p_t = test[test['promocode']==1]
       p_t
```

t[184]:

visitid	promocode	eVar82	order
100010825755995377356241128984433946298 1	True	False	True
100034118541581008486617653558704628597	True	False	True
100068898285134821927141456802434564309	True	False	True
100082663407957129016555099531533098 91	True	False	True
100082663407957129016555099531533098 92	True	False	True
...	...	...	...
998989327524520500336141267982577500 96	True	False	True
998989327524520500336141267982577500 97	True	False	True
999668305393118804481909471437022066 52	True	False	True
999819099848648206449466642371149422 71	True	False	True
999865677449219052734349735519793826 21	True	False	False

55284 rows × 3 columns



Back Up Slides

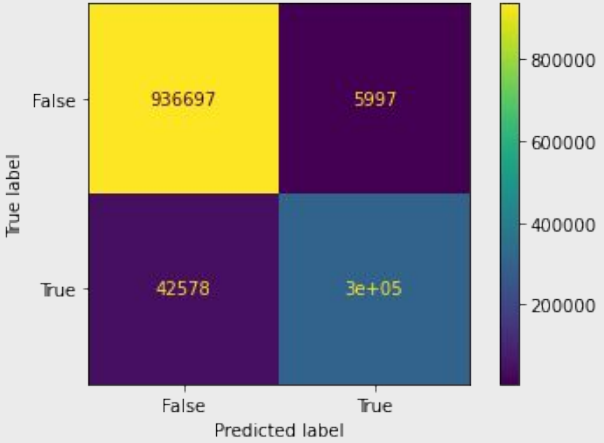
Variable Importance for RF

1. Mean Decrease in Impurity (MDI)

- Same as Bagging.
- Calculate the total amount that the MSE (for regression) or Gini index (for classification) is decreased due to splits over a given predictor in different nodes (weighted by the probability of reaching that node (which is approximated by the proportion of samples reaching that node)).
- The weighted decrease in purity as a result of the splits over a given predictor is averaged over all trees, and is used as a measure of the importance of variable j in the random forest.
- The default in Scikit-learn `feature_importances_`

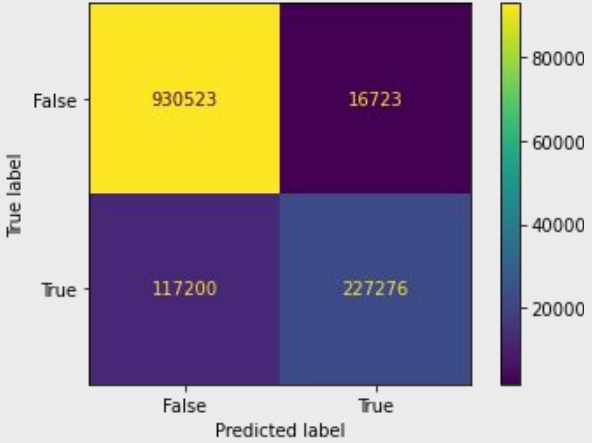
Random Forest

F1-Score:	0.9256



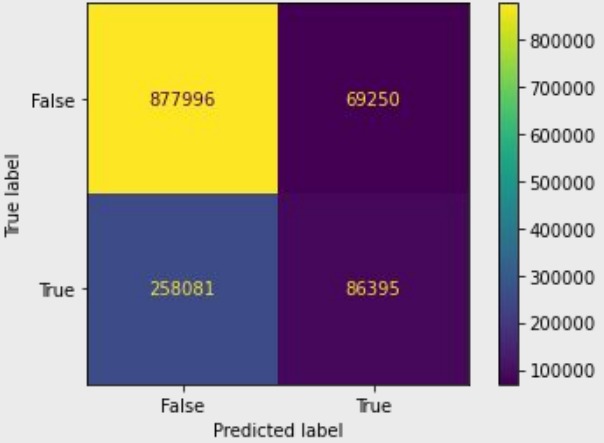
Gradient Boosting

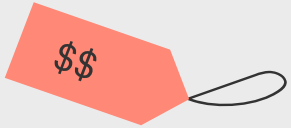
F1-Score:	0.7724



Logistic Regression

F1-Score:	0.3454





Resources

Slide template: www.slidesgo.com