Berkeley Engineering
BerkeleyHaas

Shopping Coupon Recommendation using Ensemble Models

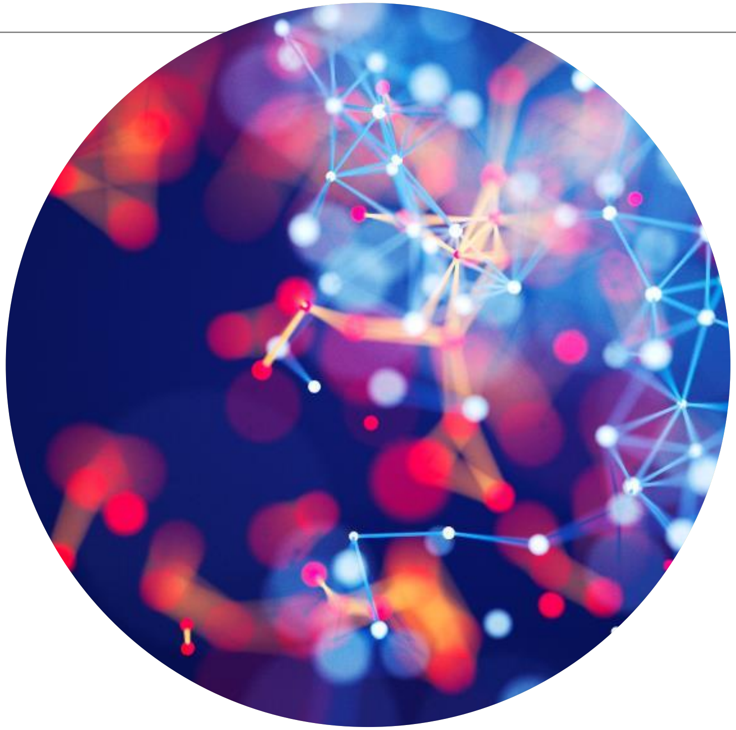
VINITHA JEEVARATHNAM

August 2022

Professional Certificate in Machine Learning and Artificial Intelligence – Capstone Project

Agenda

- Introduction
- Business Objective
- Exploratory Data Analysis (EDA)
- Ensemble Modeling
- Conclusion
- Explainer Dashboard
- Future Recommendations



Introduction to Recommendation Systems

Benefits

- Increase in Conversion Rate
- Increases chances of upselling
- Customer Retention & Loyalty
- Customer Satisfaction
- Overall Revenue Growth

- ❑ Recommendation models play an important role in the success of e-commerce business
- ❑ Tailoring shopping recommendations to consumers and merchants assist in making a purchase decision
- ❑ Shopping personalization uses consumers past behavior to predict their future needs and offer products or services recommendations accordingly
- ❑ Personalization recommendation is increasingly becoming a must for consumers than a nice to have option
- ❑ McKinsey & Company 2021 article on personalization




Business Objective

In this capstone project, a customer ecommerce shopping transaction data from Kaggle is used to evaluate data and compare the performance of recommendations models that are built based on Ensemble Modeling concept. The business objective of this project is to identify the feature (like brand, product etc.) that influences the user to accept or reject coupons.

Exploratory Data Analysis

Dataset and Data Attributes

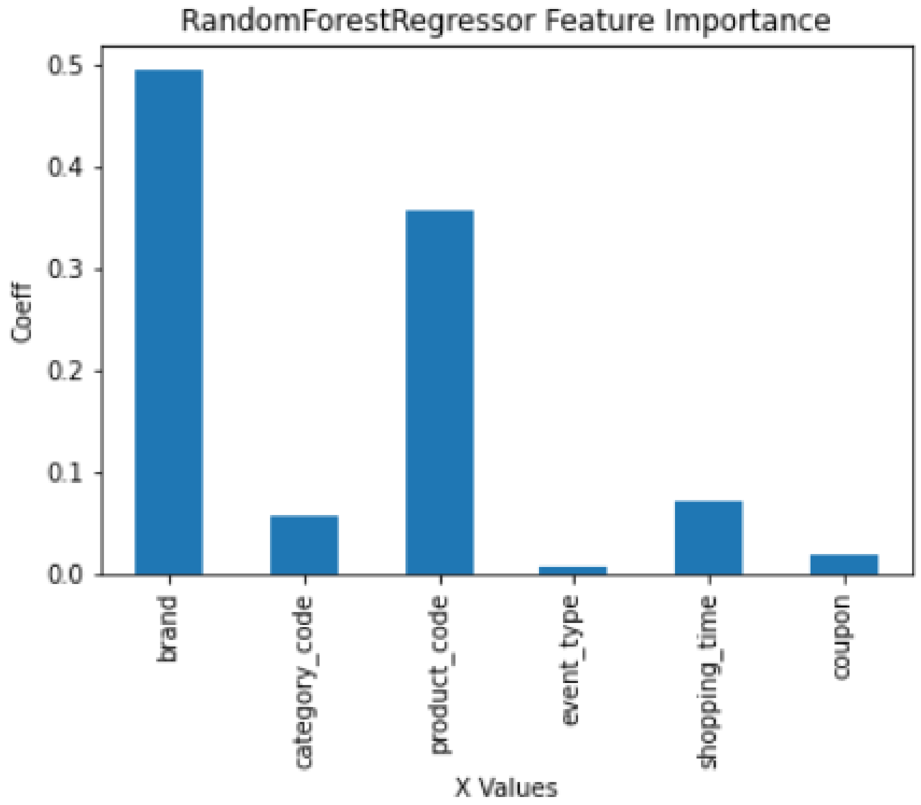
- **Dataset** (from [Kaggle]): A large multi-category eCommerce store user behavior data for one month (November 2019). Contains 42,448,764 records.
- Each row in the file represents a user event. All events are related to products and users. There are different types of events.
  - **event_time:** Time when event happened at (in UTC) – Converted to **shopping_time**
  - **event_type:** (Typical funnel: view => cart => purchase)
    - view - a user viewed a product
    - cart - a user added a product to shopping cart
    - purchase - a user purchased a product
  - **product_id:** ID of a product – Dropped
  - **category_id:** Product's category ID – Dropped
  - **category_code:** multi-hierarchical name of Product's category
  - **brand:** Downcased string of brand name
  - **price:** Float price of a product
  - **user_id:** Permanent user ID – Dropped
  - **user_session:** Temporary user's session ID. Same for each user's session. Is changed every time user come back to online store from a long pause – Dropped
  - **Coupon:** synthetically generated feature to show whether the user used a coupon or not (with flags Yes or No)

EDA Outcomes

- Since the dataset is too large, stratified the data to create a sample of 53,121 records
- The *‘event_type’* column in time-series format is converted to *‘shopping_time’*
- Multilevel ‘category’ values are converted to single value (level 3)
- Used *corr()* and identified ‘price’ and *‘shopping_time’* has higher correlation
- Used *seaborn* and *plotly* charts to visualize data and inferred the following,
  - ***Electronics*** is the top shopping category
  - ***Smartphone*** is the top shopping product
  - ***Apple*** and ***Samsung*** are the top shopping brands
  - ***View***-ing products is the top *‘event_type’* and most *‘shopping_time’* spent user action
    - Acceptance of *‘coupon’* is higher for this event type
  - Acceptance of *‘coupon’* is NOT influenced by *‘shopping_time’* spent by a user
  - Users spent more *‘shopping_time’* on electronics, auto, and appliances shopping category

Feature Importance/Selection using RandomForestRegressor

- Used **‘price’** as the regressor target feature
- Among the categorical features, ‘coupon’ exhibits higher feature importance than ‘event_type’
- For modeling, **‘coupon’** is used as the target feature

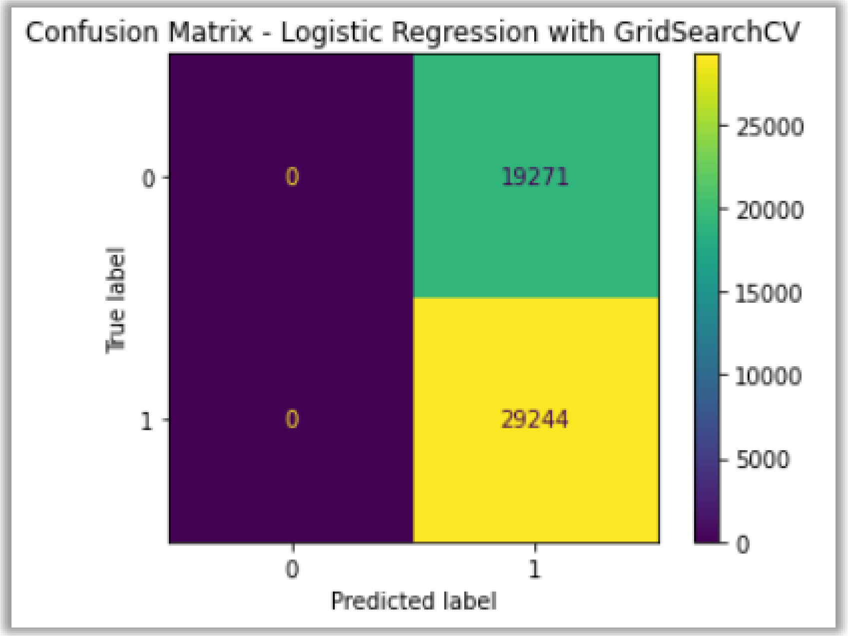


LogisticRegression

- Baseline Model
- With GridSearchCV

Logistic Regression with GridSearchCV

	Metrics	Results
AdaBoost	**Mean Absolute Error (MAE)**	0.39722
	Cross Val Score	0.60278
	R2 Score	-0.65897
	TP, TN, FP, FN	0, 29244, 19271, 0
AdaBoost with GridSearchCV	**Mean Absolute Error (MAE)**	0.39722
	Cross Val Score	0.60276
	R2 Score	-0.65897
	Mean Fit Time	0.13806
	Best Params	{'max_iter': 30}
	TP, TN, FP, FN	0, 29244, 19271, 0

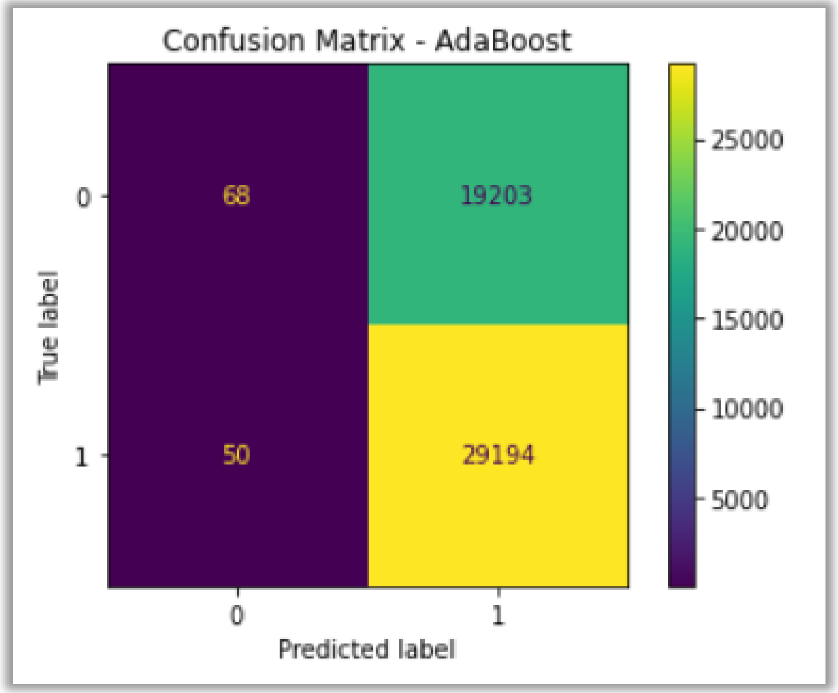


Ensemble Models - Boosting Classifiers

- AdaBoostClassifier
- GradientBoostingClassifier
- XGBClassifier
- HistGradientBoostingClassifier

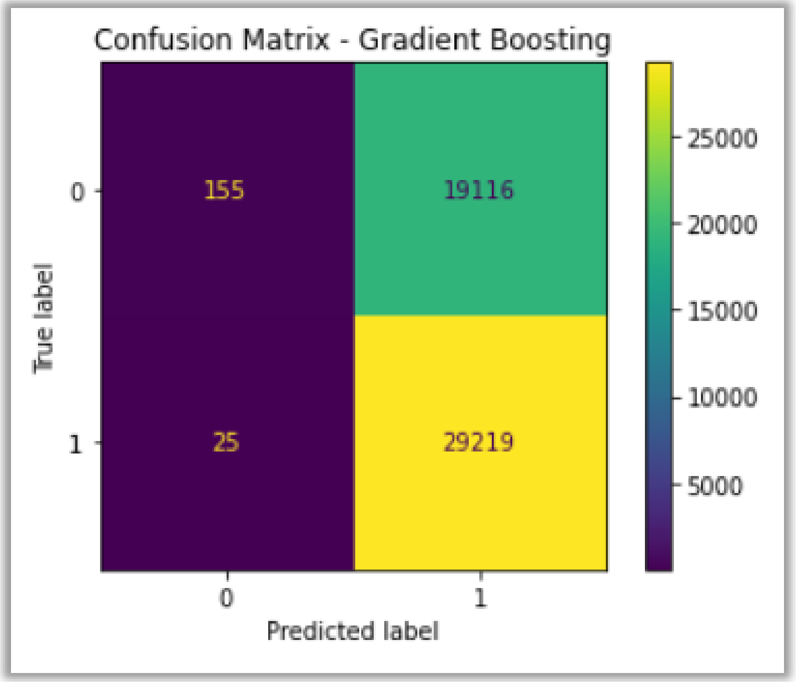
AdaBoostClassifier with GridSearchCV

	Metrics	Results
AdaBoost	Mean Absolute Error (MAE)	0.39685
	Cross Val Score	0.60206
	R2 Score	-0.65742
	TP, TN, FP, FN	68, 29194, 19203, 50
AdaBoost with GridSearchCV	Mean Absolute Error (MAE)	0.39722
	Cross Val Score	0.60192
	R2 Score	-0.65897
	Mean Fit Time	0.99839
	Best Params	{'n_estimators': 30}
	TP, TN, FP, FN	3, 29241, 19268, 3



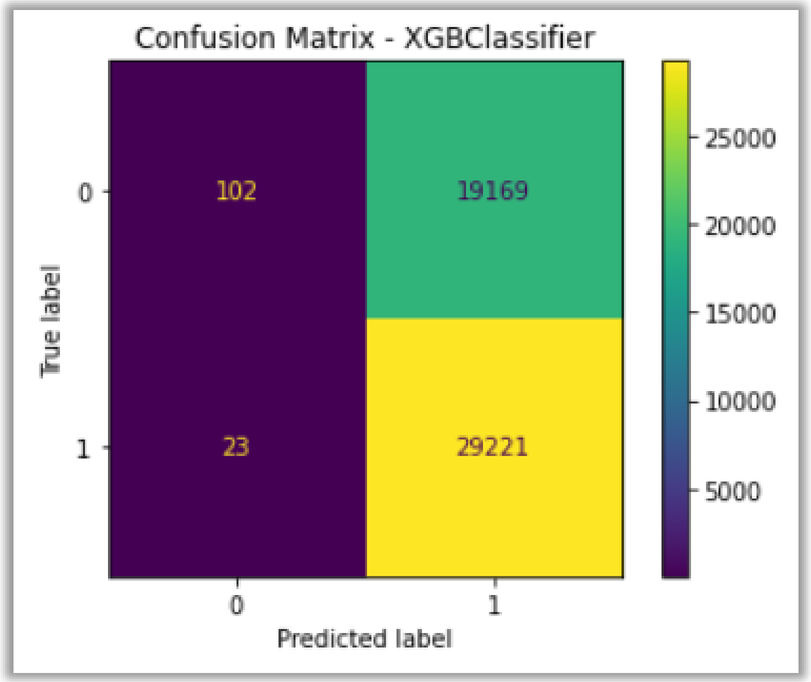
GradientBoostingClassifier with GridSearchCV

	Metrics	Results
Gradient Boosting	**Mean Absolute Error (MAE)**	0.39454
	Cross Val Score	0.60163
	R2 Score	-0.64778
	TP, TN, FP, FN	155, 29219, 19116, 25
Gradient Boosting with GridSearchCV	**Mean Absolute Error (MAE)**	0.39705
	Cross Val Score	0.60264
	R2 Score	-0.65828
	Mean Fit Time	1.47587
	Best Params	{'n_estimators': 10}
	TP, TN, FP, FN	8, 29244, 19263, 0



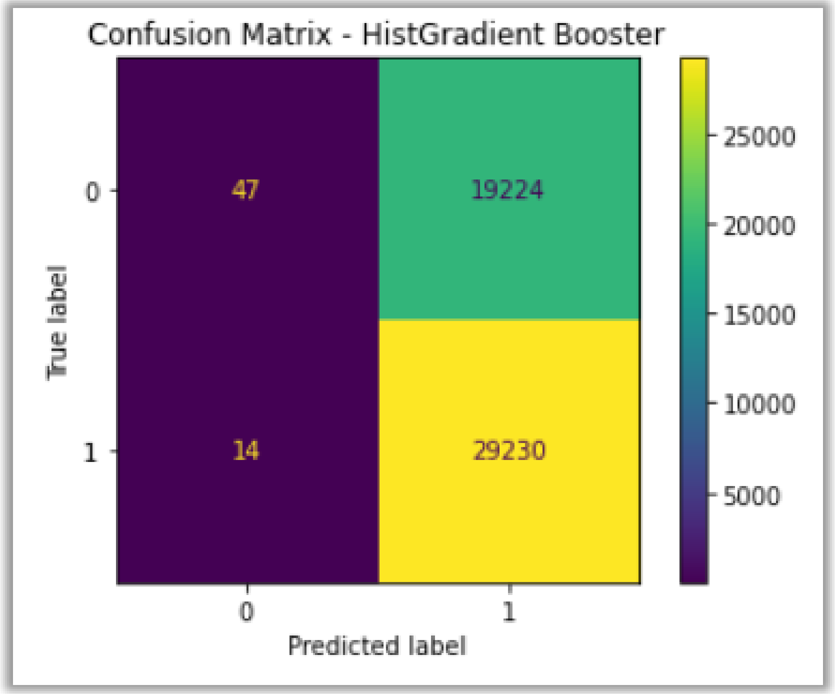
XGBClassifier with GridSearchCV

	Metrics	Results
XGBClassifier	**Mean Absolute Error (MAE)**	0.39559
	Cross Val Score	0.60233
	R2 Score	-0.65217
	TP, TN, FP, FN	102, 29221, 19169, 23
XGBClassifier with GridSearchCV	**Mean Absolute Error (MAE)**	0.39648
	Cross Val Score	0.60235
	R2 Score	-0.65587
	Mean Fit Time	0.70548
	Best Params	{'n_estimators’: 10}
	TP, TN, FP, FN	46, 29234, 19225, 10



HistGradientBoostingClassifier with GridSearchCV

	Metrics	Results
HistGradientBoosting	**Mean Absolute Error (MAE)**	0.39654
	Cross Val Score	0.60256
	R2 Score	-0.65613
	TP, TN, FP, FN	47, 29230, 19224, 14
HistGradientBoosting with GridSearchCV	**Mean Absolute Error (MAE)**	0.39654
	Cross Val Score	0.60276
	R2 Score	-0.65613
	Mean Fit Time	0.17934
	Best Params	{‘max_iter’: 10}
	TP, TN, FP, FN	47, 29230, 19224, 14



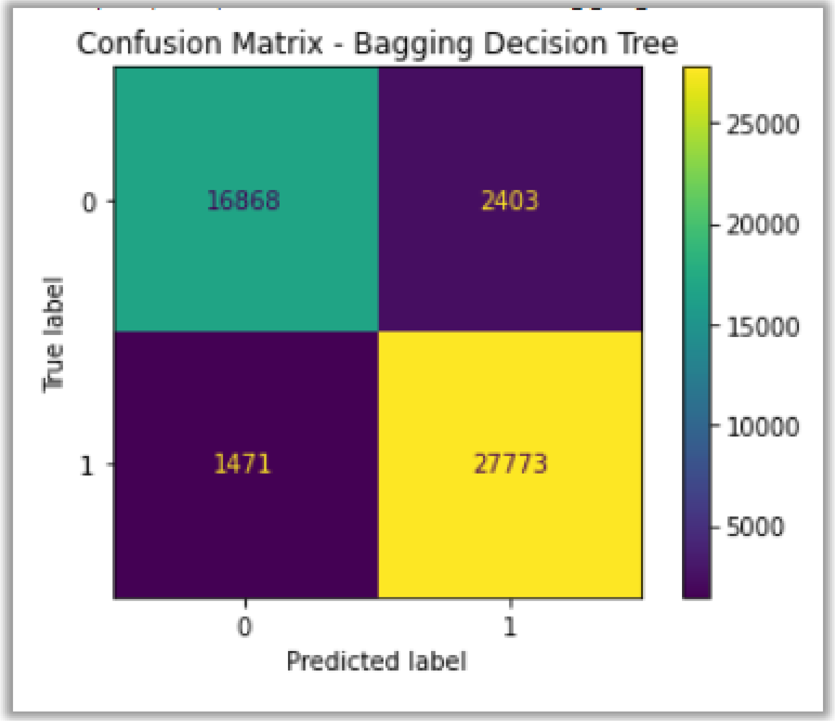
Ensemble Models - Bagging Classifiers

- DecisionTreeClassifer
- ExtraTreeClassifier

*RandomForestClassifier and SVC – Unable to use due to limited computational resources

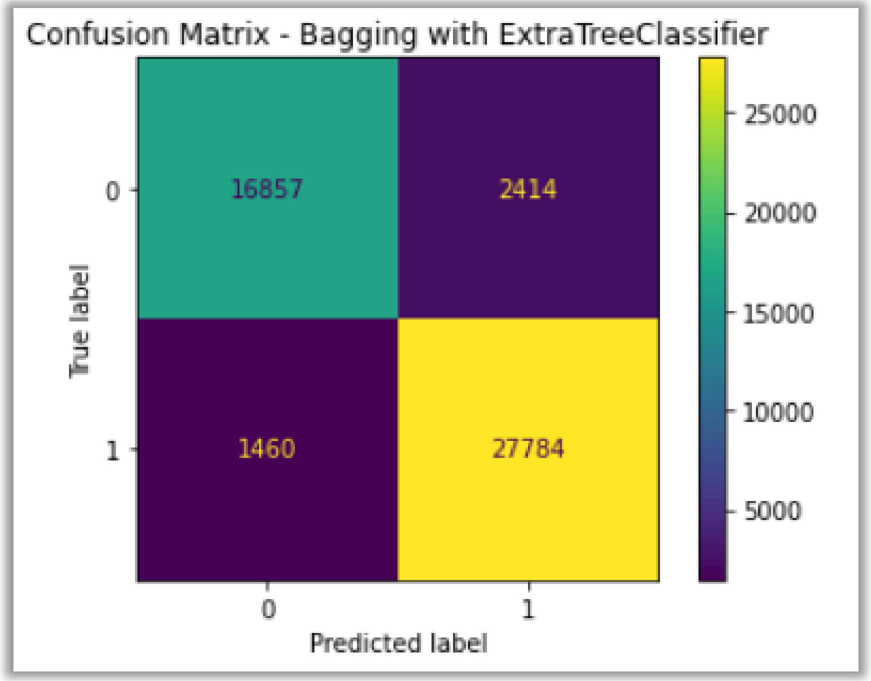
DecisionTreeClassifier with GridSearchCV

	Metrics	Results
DecisionTreeCl assifier	Mean Absolute Error (MAE)	0.07985
	Cross Val Score	0.53311
	R2 Score	0.66650
	TP, TN, FP, FN	16868, 27773, 2403, 1471
DecisioTreeClassifier with GridSearchCV	Mean Absolute Error (MAE)	0.07985
	Cross Val Score	0.53287
	R2 Score	0.66650
	Mean Fit Time	5.83084
	Best Params	{‘n_estimators’: 100}
	TP, TN, FP, FN	16868, 27773, 2403, 1471



ExtraTreeClassifier with GridSearchCV

	Metrics	Results
ExtraTreeClassifier	Mean Absolute Error (MAE)	0.07985
	Cross Val Score	0.53060
	R2 Score	0.66650
	TP, TN, FP, FN	16857, 27784, 2414, 1460
ExtraTreeClassifier with GridSearchCV	Mean Absolute Error (MAE)	0.07985
	Cross Val Score	0.53041
	R2 Score	0.66650
	Mean Fit Time	1.09442
	Best Params	{'n_estimators': 100}
	TP, TN, FP, FN	16857, 27784, 2414, 1460

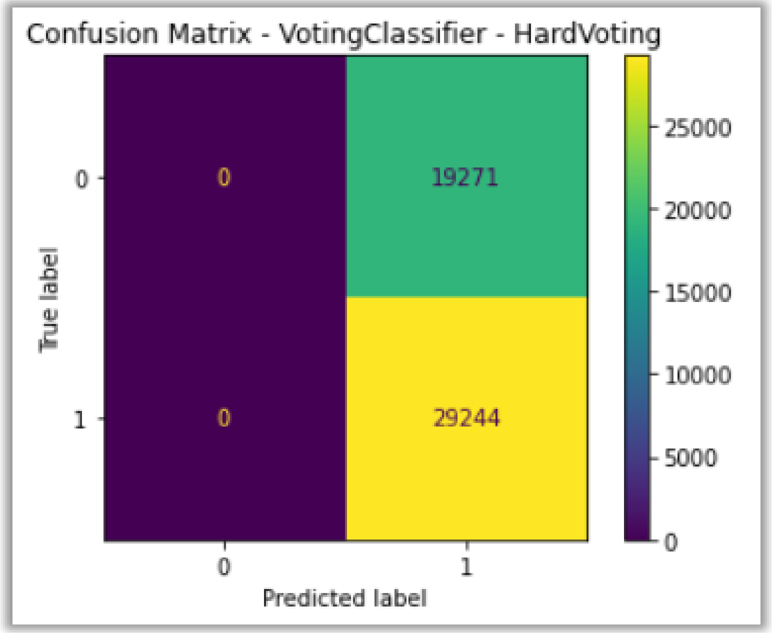


Ensemble Models – VotingClassifier

- ❑ HardVoting
- ❑ SoftVoting

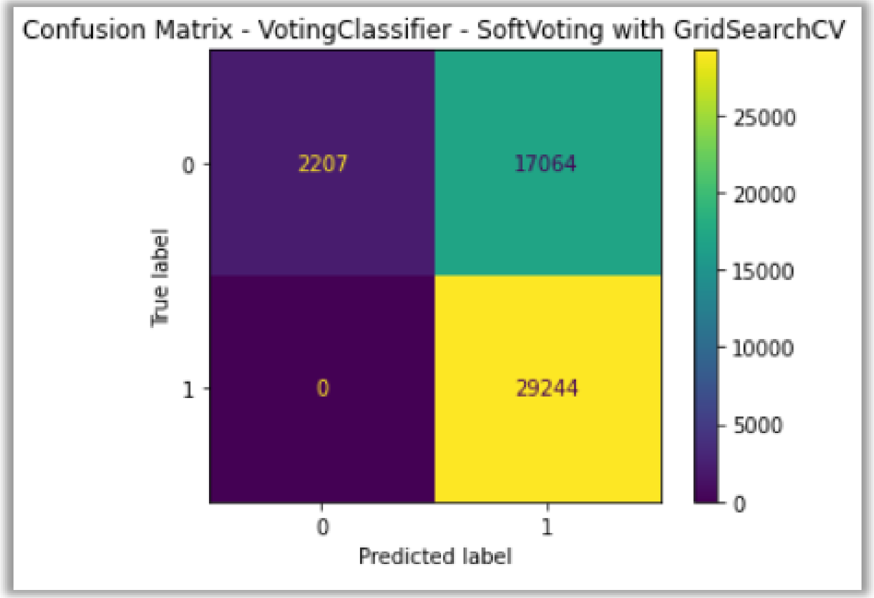
VotingClassifier-HardVoting with GridSearchCV

	Metrics	Results
VotingClassifier - HardVoting with GridSearchCV	**Mean Absolute Error (MAE)**	0.39722
	Cross Val Score	0.60278
	R2 Score	-0.65897
	Mean Fit Time	2.54453
	Best Params	{'lr__C': 1.0, 'rf__n_estimators': 10}
	TP, TN, FP, FN	0, 29244, 19271, 0



VotingClassifier-SoftVoting with GridSearchCV

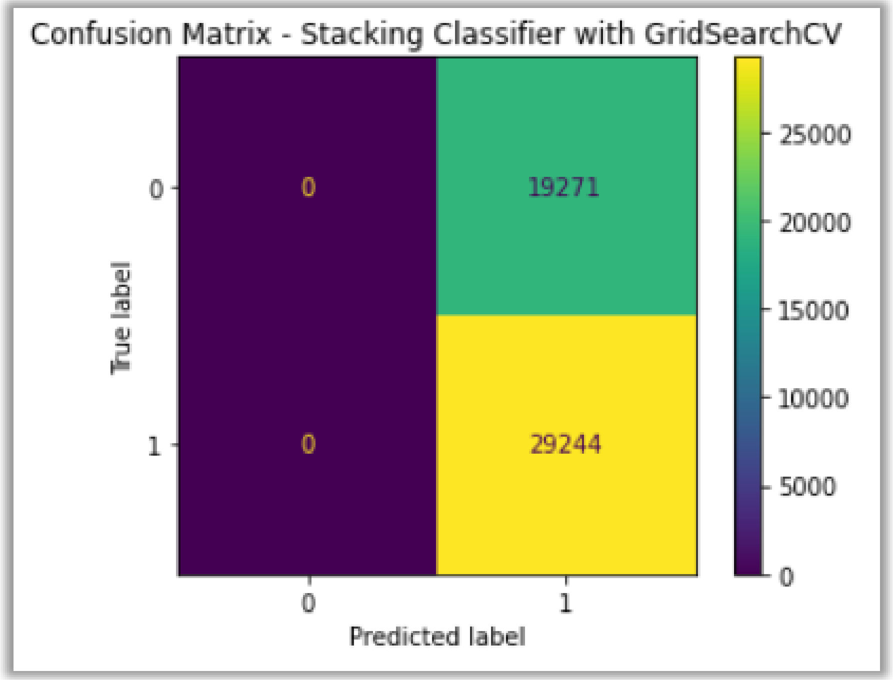
	Metrics	Results
VotingClassifier - HardVoting with GridSearchCV	**Mean Absolute Error (MAE)**	0.3517262702257034
	Cross Val Score	0.5979800061836545
	R2 Score	-0.46897989365213943
	Mean Fit Time	5.701870036125182
	Best Params	{'lr__C': 1.0, 'rf__n_estimators': 200}
	TP, TN, FP, FN	2207, 29244, 17064, 0



Ensemble Models – StackingClassifier

StackingClassifier with GridSearchCV

	Metrics	Results
StackingClassifier with GridSearchCV	**Mean Absolute Error (MAE)**	0.39721735545707515
	Cross Val Score	0.602782644542925
	R2 Score	-0.658972780741349
	Mean Fit Time	3.0968246459960938
	Best Params	{'rf__n_estimators': 5}
	TP, TN, FP, FN	0, 29244, 19271, 0



Conclusion

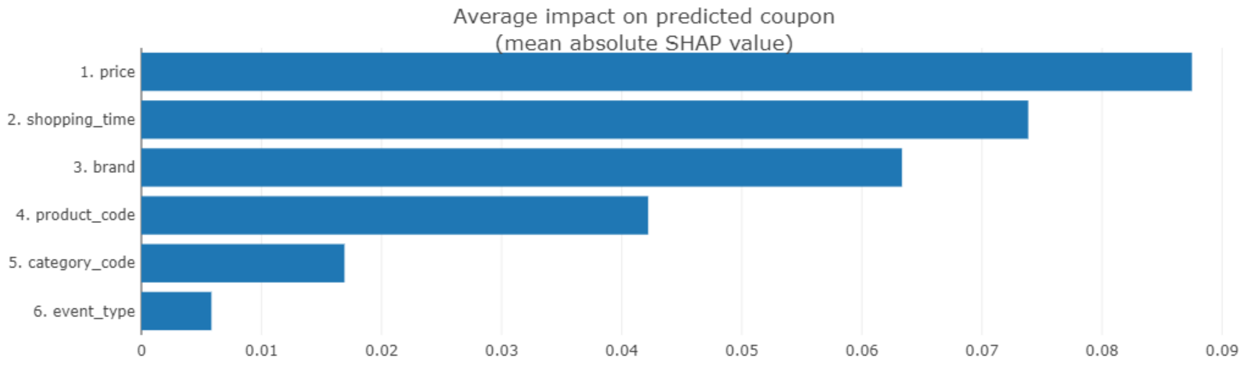
In this capstone project, we have evaluated LogisticRegression and 9 Ensemble Models using Boosting, Bagging, Voting and Stacking Classifiers. Based on the metrics collected, the following models performed well,

1. Bagging DecisionTreeClassifier WITH/WITHOUT GridSearchCV
2. Bagging ExtraTeeClassifier WITH/WITHOUT GridSearchCV
3. Soft VotingClassifer WITH GridSearchCV

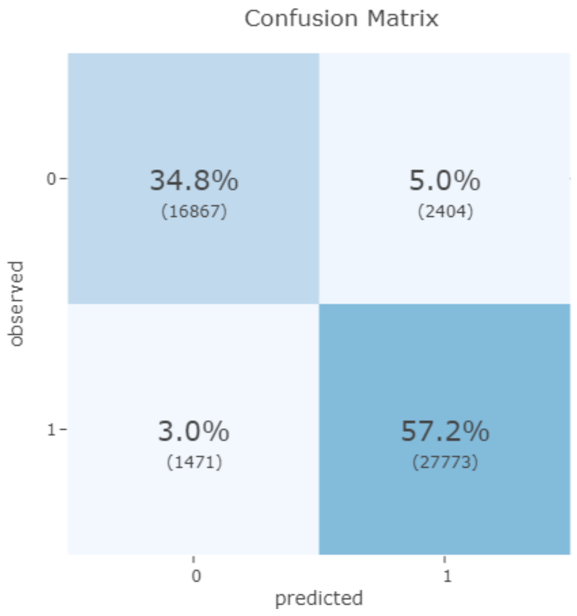
Based on the other metrics like fit_time, FP, FN, etc, we can conclude that option 1 is the appropriate model to select based on the dataset and project business objective.

Explainer Dashboard

Feature Importance



Confusion Matrix



Model Performance Metrics

Metric	Score
accuracy	0.92
precision	0.92
recall	0.95

Metric	Score
f1	0.935
roc_auc_score	0.983
pr_auc_score	0.989
log_loss	0.273

*For detailed list of Explainer Dashboard is available here

Future Recommendations

- The ensemble models can be applied to *'event_type'* to understand whether the model accuracy improves
- Adding more data around user profile like age, sex, education, job etc. will further improve the targeting of coupons to users
- Adding Merchant Catalog or Coupon information will further enhance the recommendation of coupons to the users as well as improves revenue generation for the merchants

Thank You

Vinitha Jeevarathnam

vinitha.Jeevarathnam@gmail.com

GitHub: https://github.com/vinithajeeva/ai-ml-capstone-project