FUNDAMENTALS OF DATA SCIENCE
WINTER 2023-24

FINAL PROJECT PRESENTATION – DEC 18TH, 2023

ANALYZING KOBE BRYANT'S SHOT SUCCESS

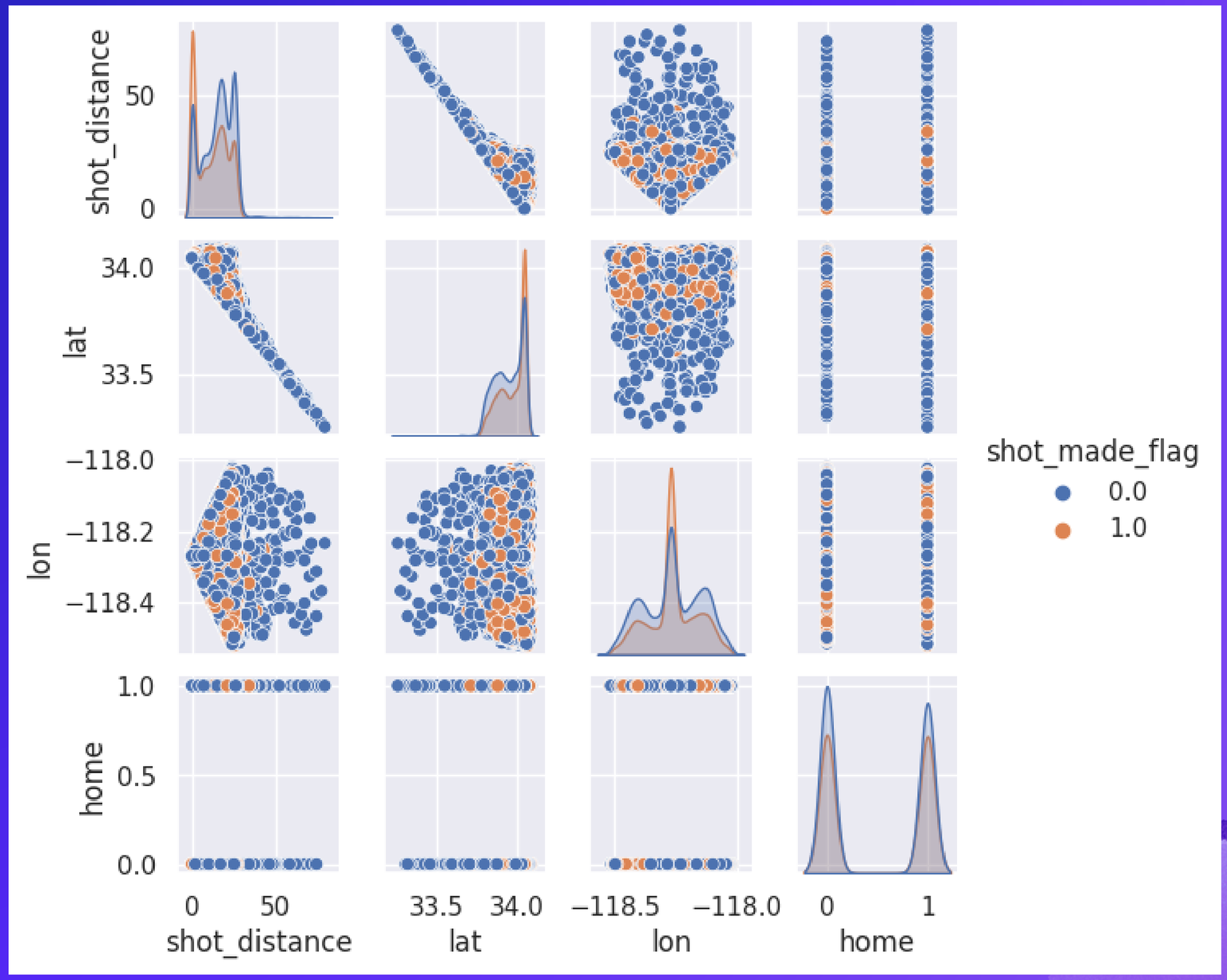
E. MOKHTARI, A. BAKHSHAEE, S. SEYYEDI PARSA, O. GHORBANI

SHEMATIC OF THE DATA

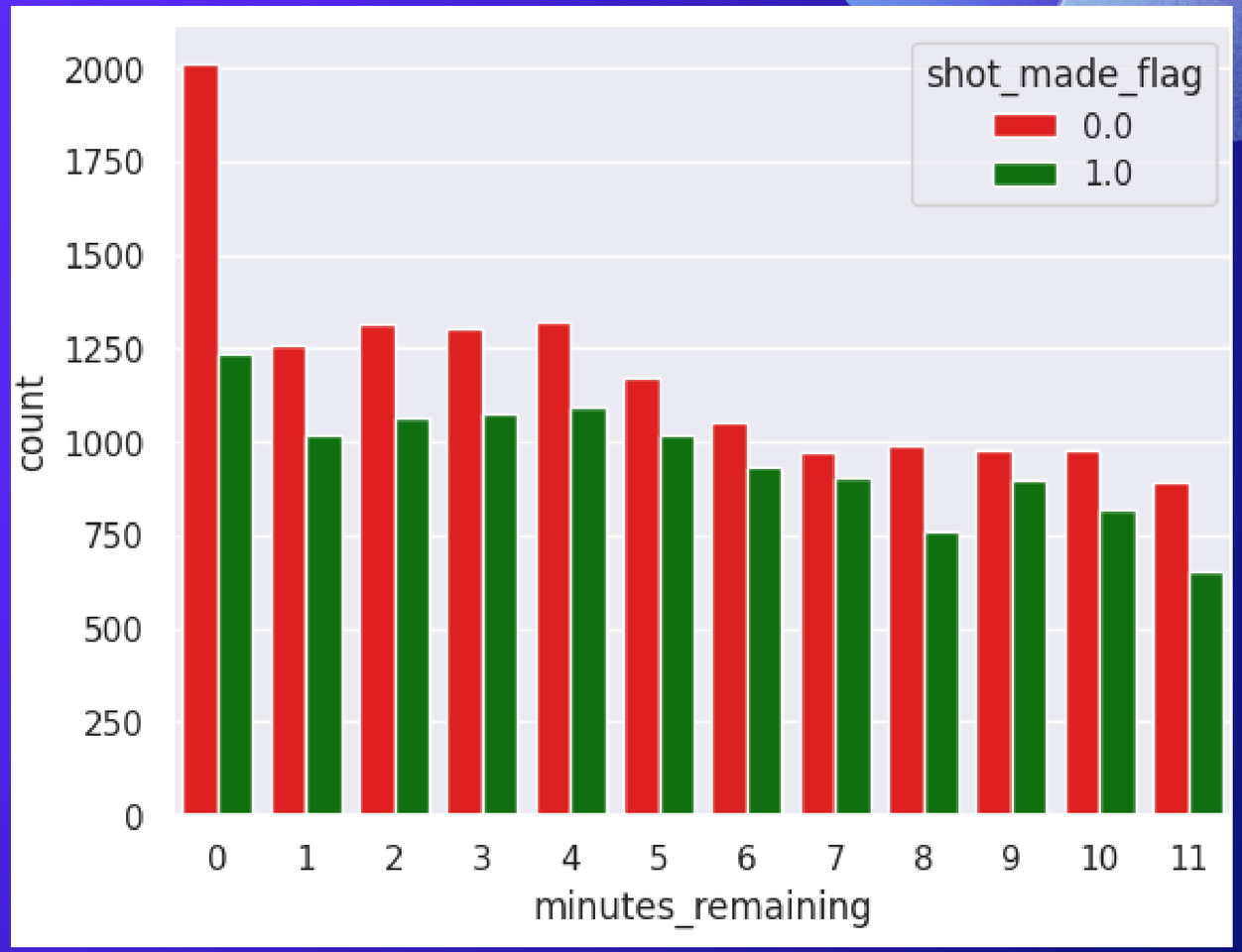
action_type	combined_shot_type	game_event_id	game_id	lat	loc_x	loc_y	lon	minutes_remaining	period	...	shot_type	shot_zone_area
Jump Shot	Jump Shot	10	20000012	33.9723	167	72	-118.1028	10	1	...	2PT Field Goal	Right Side(R)
Jump Shot	Jump Shot	12	20000012	34.0443	-157	0	-118.4268	10	1	...	2PT Field Goal	Left Side(L)
Jump Shot	Jump Shot	35	20000012	33.9093	-101	135	-118.3708	7	1	...	2PT Field Goal	Left Side Center(LC)
Jump Shot	Jump Shot	43	20000012	33.8693	138	175	-118.1318	6	1	...	2PT Field Goal	Right Side Center(RC)
Driving Dunk Shot	Dunk	155	20000012	34.0443	0	0	-118.2698	6	2	...	2PT Field Goal	Center(C)

```
Index(['action_type', 'combined_shot_type', 'game_event_id', 'game_id', 'lat',
       'loc_x', 'loc_y', 'lon', 'minutes_remaining', 'period', 'playoffs',
       'season', 'seconds_remaining', 'shot_distance', 'shot_made_flag',
       'shot_type', 'shot_zone_area', 'shot_zone_basic', 'shot_zone_range',
       'team_id', 'team_name', 'game_date', 'matchup', 'opponent', 'shot_id'],
      dtype='object')
Number of the columns at first :  25
```

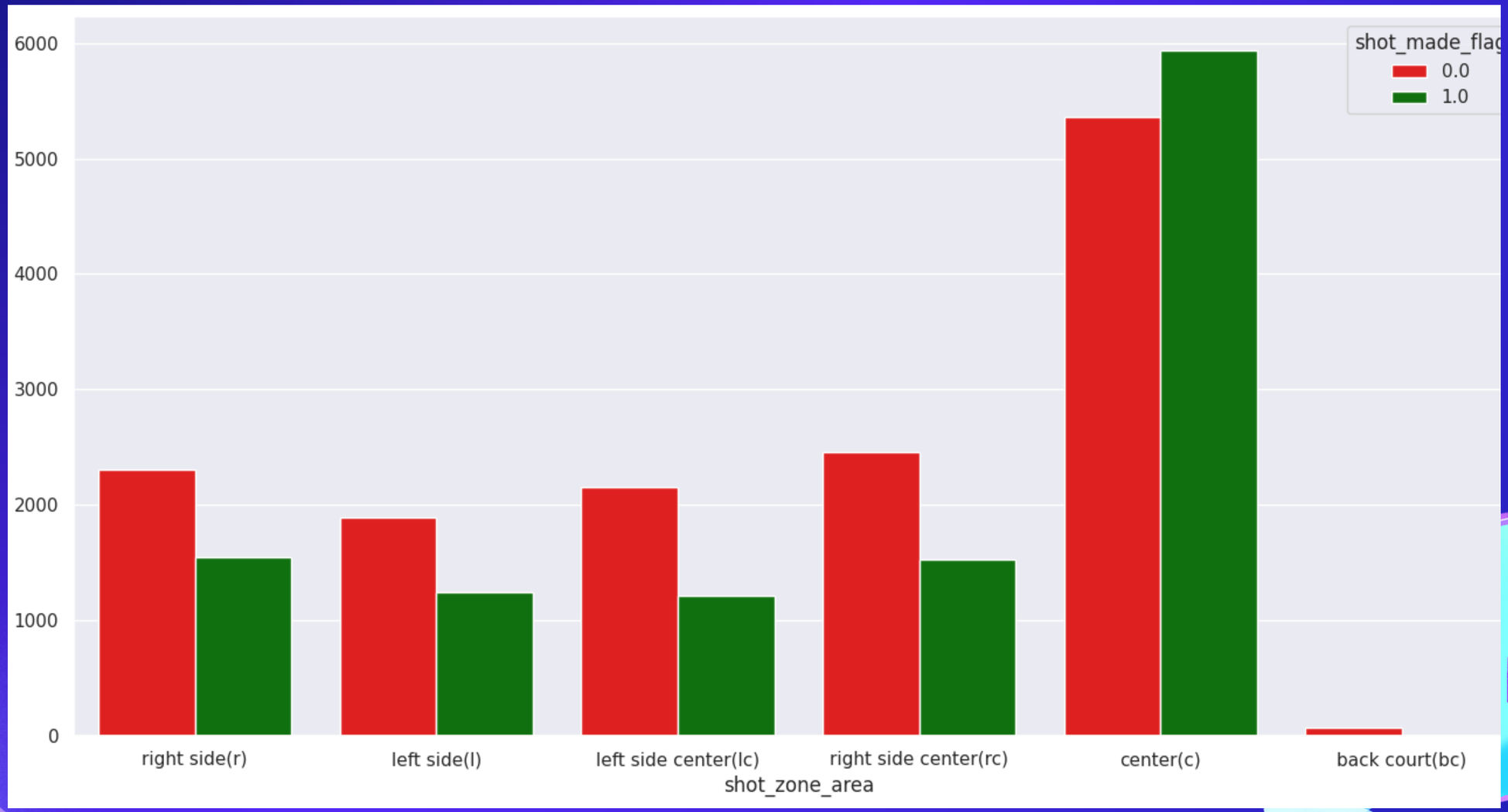
Data visulization at first glance (EDA)



Data visulization at first glance (EDA)



Data visulization at first glance (EDA)



CLEANING AND PREPROCESSING THE DATA

01

216 features

02

Removing missing values and duplicate

03

Handling categorical feature using ordinal and one hot encoding

```
from sklearn.preprocessing import OrdinalEncoder

distance_order = ['less than 8 ft.', '8-16 ft.', '16-24 ft.', '24+ ft.', 'back court shot']

# Initialize the OrdinalEncoder with the specified order
ordinal_encoder = OrdinalEncoder(categories=[distance_order], dtype=int)

new = ordinal_encoder.fit_transform(df[['shot_zone_range']])

df['shot_zone_range'] = new
```

FEATURE SELECTION

CHI2 FEATURE SELECTION

RECURSIVE FEATURE ELIMINATION

VARIANCE THRESHOLD

Reduce the dimention of data into 48 features

Handle Imbalance data usign SMOTE algorithm

ALGORITHMS DEVELOPED

LOGISTIC REGRESSION TRAINING ACCURACY: 0.7048

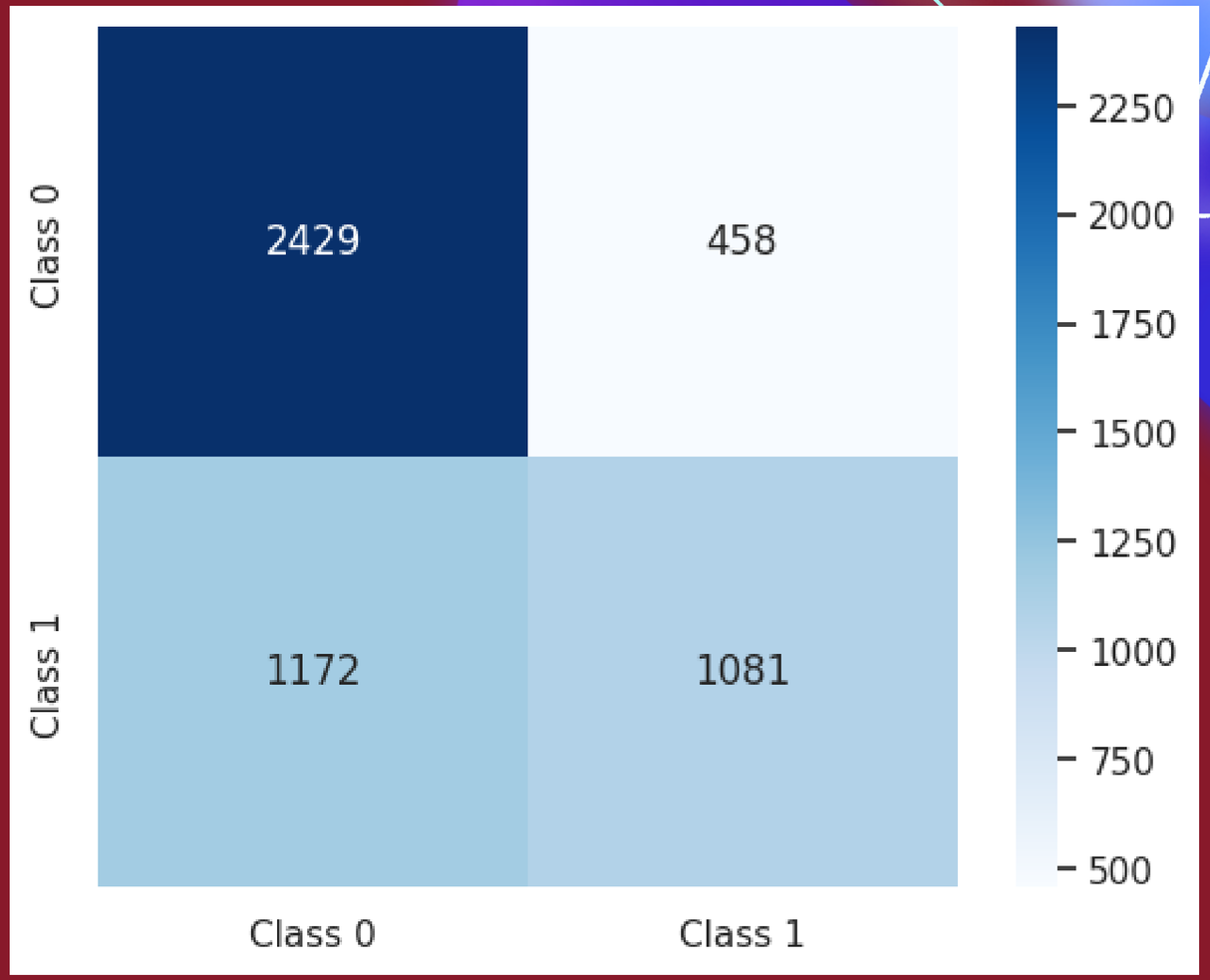
GRADIENT BOOSTING CLASSIFIER TEST ACCURACY: 0.6804

RANDOM FOREST CLASSIFIER TEST ACCURACY: 0.6825

VOTING ENSEMBLE TEST ACCURACY 0.6829

FINAL RESULT

VOTING ENSEMBLE
TEST ACCURACY: 0.6829




```
from sklearn.ensemble import BaggingClassifier, ExtraTreesClassifier, GradientBoostingClassifier, VotingClassifier, RandomForestClassifier, AdaBoostClassifier

# Create sub models
estimators = []

estimators.append(('gbm', GradientBoostingClassifier(n_estimators=200, max_depth=3, learning_rate=0.1, max_features=15, warm_start=True, random_state=42)))
estimators.append(('rf', RandomForestClassifier(bootstrap=True, max_depth=8, n_estimators=200, max_features=20, criterion='entropy', random_state=42)))
estimators.append(('ada', AdaBoostClassifier(algorithm='SAMME.R', learning_rate=1e-2, n_estimators=10, random_state=42)))

# create the ensemble model
ensemble = VotingClassifier(estimators, voting='soft', weights=[3,3,1])
```

THANK YOU