

```
In [1]: import numpy as np # Linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import seaborn as sns
import itertools
import matplotlib.pyplot as plt
import string
import re
import collections
from sklearn import preprocessing
from wordcloud import WordCloud
from sklearn.model_selection import train_test_split, KFold, cross_val_score
from xgboost import XGBClassifier
import xgboost as xgb
from sklearn.metrics import make_scorer, f1_score, accuracy_score, mean_absolute_err
import optuna
from lofo import LOFOImportance, Dataset, plot_importance
%matplotlib inline
import itertools
```

```
In [2]: # READ DATA
train_df = pd.read_json('train.json.zip')
test_df = pd.read_json('test.json.zip')
```

```
In [3]: train_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 49352 entries, 4 to 124009
Data columns (total 15 columns):
#   Column                Non-Null Count  Dtype
---  -
0   bathrooms              49352 non-null  float64
1   bedrooms               49352 non-null  int64
2   building_id            49352 non-null  object
3   created                49352 non-null  object
4   description             49352 non-null  object
5   display_address        49352 non-null  object
6   features               49352 non-null  object
7   latitude               49352 non-null  float64
8   listing_id             49352 non-null  int64
9   longitude              49352 non-null  float64
10  manager_id             49352 non-null  object
11  photos                 49352 non-null  object
12  price                  49352 non-null  int64
13  street_address         49352 non-null  object
14  interest_level         49352 non-null  object
dtypes: float64(3), int64(3), object(9)
memory usage: 6.0+ MB
```

\*\*Our target 'INTEREST LEVEL' is an object as we can see above.

\*\*Let's convert to the numeric to analyze easily

- 0 : low
- 1 : medium
- 2 : high

```
In [4]: train_df['target'] = train_df['interest_level'].apply(lambda x: 0 if x=='low'
                                                             else 1 if x=='medium'
                                                             else 2)
# train_df['low'] = train_df['interest_level'].apply(lambda x: 1 if x=='low' else 0)
```

```
# train_df['medium'] = train_df['interest_level'].apply(lambda x: 1 if x=='medium' else 0)
# train_df['high'] = train_df['interest_level'].apply(lambda x: 1 if x=='high' else 0)
```

## BASIC FEATURES

```
In [5]: train_df['description'].iloc[0]
```

```
Out[5]: 'Spacious 1 Bedroom 1 Bathroom in Williamsburg!Apartment Features:- Renovated Eat in
Kitchen With Dishwasher- Renovated Bathroom- Beautiful Hardwood Floors- Lots of Sunl
ight- Great Closet Space- Freshly Painted- Heat and Hot Water Included- Live in Supe
r Nearby L, J, M & G Trains !<br /><br />Contact Information:Kenneth BeakExclusive A
gentC: 064-692-8838Email: kagglemanager@renthop.com, Text or Email to schedule a pri
vate viewing!<br /><br /><br /><br /><br /><br /><br /><br /><br /><br /><br /><br /><br />
<br /><br /><br /><br /><p><a  website_redacted '
```

```
In [6]: # REMOVE UNNECESSARY WORDS FROM DESCRIPTION
train_df['description'] = train_df['description'].apply(lambda x: x.replace("<br />", ""))
train_df['description'] = train_df['description'].apply(lambda x: x.replace("br", ""))
train_df['description'] = train_df['description'].apply(lambda x: x.replace("<p><a", ""))
```

```
In [7]: print(train_df['description'].iloc[0])
```

```
Spacious 1 Bedroom 1 Bathroom in Williamsburg!Apartment Features:- Renovated Eat in
Kitchen With Dishwasher- Renovated Bathroom- Beautiful Hardwood Floors- Lots of Sunl
ight- Great Closet Space- Freshly Painted- Heat and Hot Water Included- Live in Supe
r Nearby L, J, M & G Trains !Contact Information:Kenneth BeakExclusive AgentC: 064-6
92-8838Email: kagglemanager@renthop.com, Text or Email to schedule a private viewin
g!  website_redacted
```

```
In [8]: #basic features
train_df['rooms'] = train_df['bedrooms'] + train_df['bathrooms']

# count of photos #
train_df['num_photos'] = train_df['photos'].apply(len)

# count of "features" #
train_df['num_features'] = train_df['features'].apply(len)

# count of words present in description column #
train_df['num_description_words'] = train_df['description'].apply(lambda x: len(x.split()))

# description contains email
regex = r'[\w\.-]+@[\w\.-]+'
train_df['has_email'] = train_df['description'].apply(lambda x: 1 if re.findall(regex, x) else 0)

# description contains phone
# description contains phone
train_df['has_phone'] = train_df['description'].apply(lambda x: re.sub('[\d]{10}', 'phone', x).apply(
    lambda x: [s for s in x if s.isdigit()]
).apply(lambda x: len([s for s in x if len(str(s))==10]))
).apply(lambda x: 1 if x>0 else 0)

# CONVERT LOWER ALL OF WORDS
train_df[['features']] = train_df[['features']].apply(
    lambda _: [list(map(str.strip, map(str.lower, x))) for x in _])
```

## APPLY SAME OPERATIONS TO THE TEST DATA

```
In [9]: # REMOVE UNNECESSARY WORDS FROM DESCRIPTION
```

```

test_df['description'] = test_df['description'].apply(lambda x: x.replace("<br />", ""))
test_df['description'] = test_df['description'].apply(lambda x: x.replace("br", ""))
test_df['description'] = test_df['description'].apply(lambda x: x.replace("<p><a", ""))

# FEATURE ENGINEERING
#basic features
test_df['rooms'] = test_df['bedrooms'] + test_df['bathrooms']

# count of photos #
test_df["num_photos"] = test_df["photos"].apply(len)

# count of "features" #
test_df["num_features"] = test_df["features"].apply(len)

# count of words present in description column #
test_df["num_description_words"] = test_df["description"].apply(lambda x: len(x.split()))

# description contains email
regex = r'[\w\.-]+@[ \w\.-]+'
test_df['has_email'] = test_df['description'].apply(lambda x: 1 if re.findall(regex, x) else 0)

# description contains phone
test_df['has_phone'] = test_df['description'].apply(lambda x: re.sub('[^0-9]+', '', x) if len(re.sub('[^0-9]+', '', x)) > 10 else 0)

# CONVERT LOWER ALL OF WORDS
test_df[["features"]] = test_df[["features"]].apply(
    lambda _: [list(map(str.strip, map(str.lower, x))) for x in _])

```

## MOST FREQUENT FEATURES EXTRACTION

In [10]:

```

feature_value_train = train_df['features'].tolist()
feature_value_test = test_df['features'].tolist()

feature_value_train
feature_value_test

feature_lst_train = []
feature_lst_test = []

for i in range(len(feature_value_train)):
    feature_lst_train += feature_value_train[i]

for i in range(len(feature_value_test)):
    feature_lst_test += feature_value_test[i]
# print(len(feature_lst)) # all features

uniq_feature_train = list(set(feature_lst_train))
uniq_feature_test = list(set(feature_lst_test))

# print(uniq_feature) #all unique features
len(uniq_feature_train)
len(uniq_feature_test)

```

Out[10]: 1760

In [11]:

```

# see the frequency of each feature
def most_common(lst):
    features = collections.Counter(lst)

```

```

feature_value = features.keys()
frequency = features.values()
data = [('feature_value', feature_value),
        ('frequency', frequency),]
df = pd.DataFrame.from_dict(dict(data))
return df.sort_values(by = 'frequency', ascending = False)

df_features_train = most_common(feature_lst_train)
df_features_test = most_common(feature_lst_test)

df_features_train
df_features_test

```

Out[11]:

	feature_value	frequency
0	elevator	39560
8	cats allowed	35654
4	hardwood floors	35597
7	dogs allowed	33172
10	doorman	31538
...	...	...
879	skyline and river views	1
878	individual ac	1
877	** wicked w50s! * massive studio supreme * mr ...	1
876	** no broker fee! * sprawling 2br home * all b...	1
1759	** greenpoint giant! * oversized 2br masterpie...	1

1760 rows × 2 columns

In [12]:

```

def newColumn(name, df, series):
    feature = pd.Series(0,df.index,name = name)# data : 0
    for row,word in enumerate(series):
        if name in word:
            feature.iloc[row] = 1
    df[name] = feature # feature : series ; value in series : 1 or 0
    return df

# select features based on frequency
facilities = ['elevator', 'cats allowed', 'hardwood floors', 'dogs allowed', 'doorman']
for name in facilities:
    train_df = newColumn(name, train_df, train_df['features'])
    test_df = newColumn(name, test_df, test_df['features'])

```

## Features after extraction

In [13]:

```
print(train_df['features'].iloc[0])
```

```
['dining room', 'pre-war', 'laundry in building', 'dishwasher', 'hardwood floors',
'dogs allowed', 'cats allowed']
```

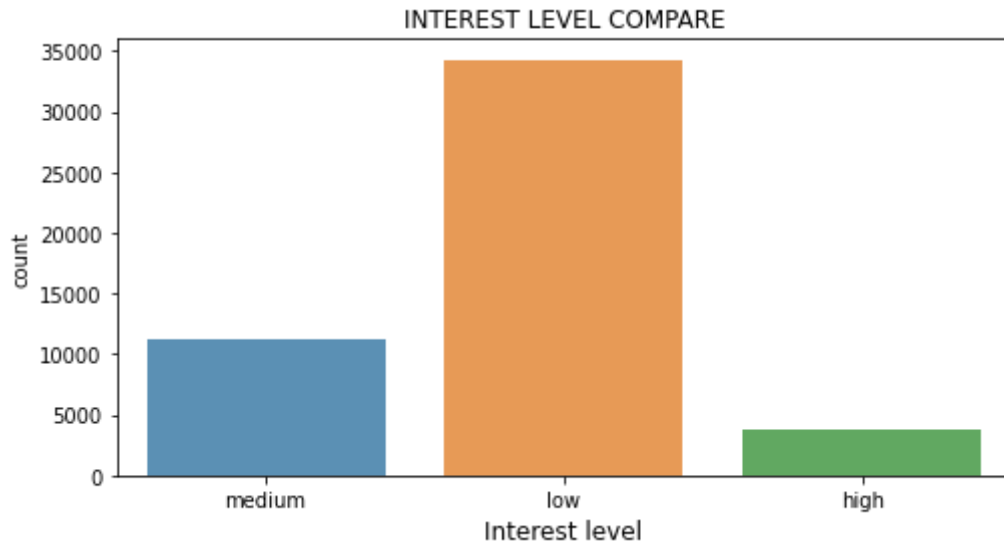
## DATA VISUALIZATION

In [14]:

```
plt.figure(figsize=(8,4))
colors = ['lightcoral','gold','lightblue']
sns.countplot(train_df['interest_level'], alpha=0.8)
plt.title("INTEREST LEVEL COMPARE")
plt.xlabel('Interest level', fontsize=12)
plt.show()
```

C:\Users\burak\AppData\Roaming\Python\Python39\site-packages\seaborn\\_decorators.py:36: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.

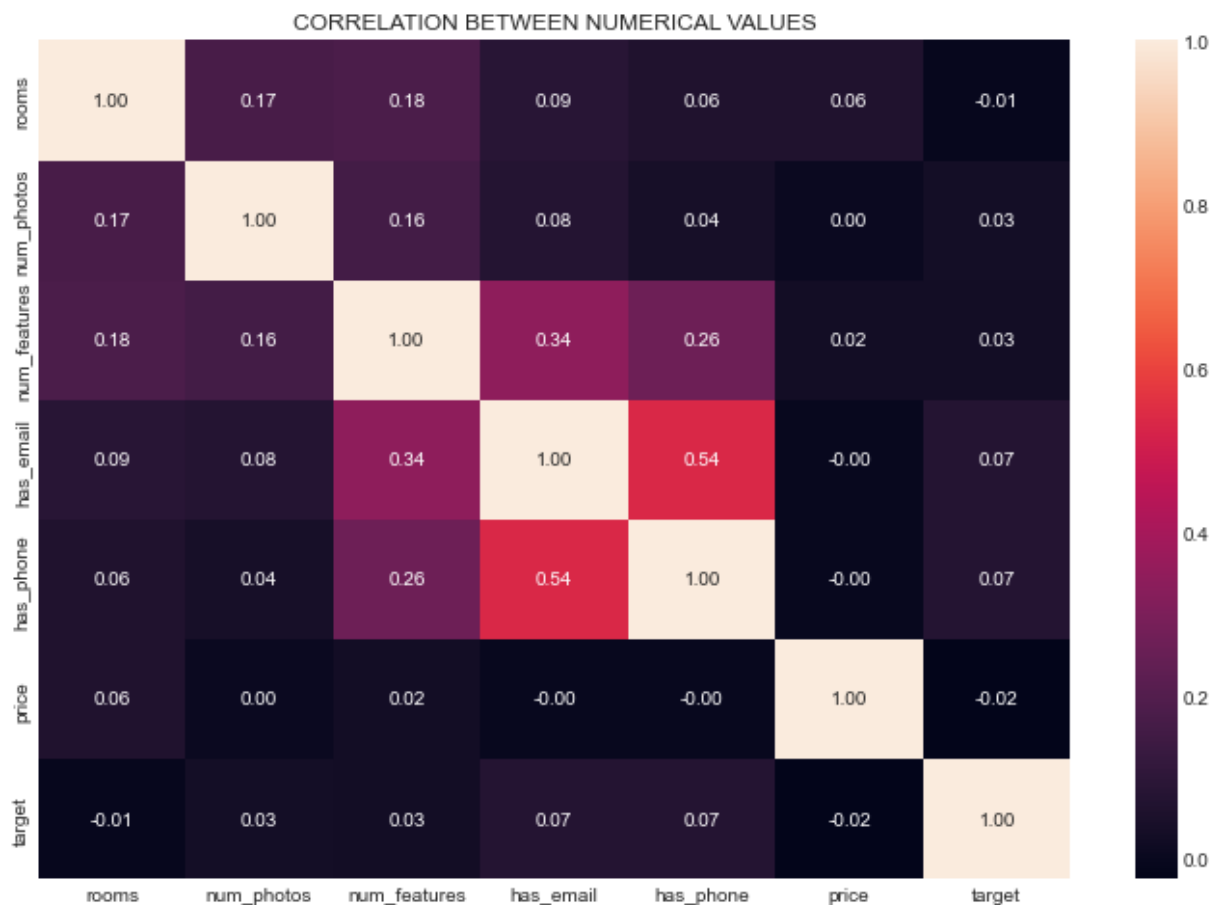
warnings.warn(



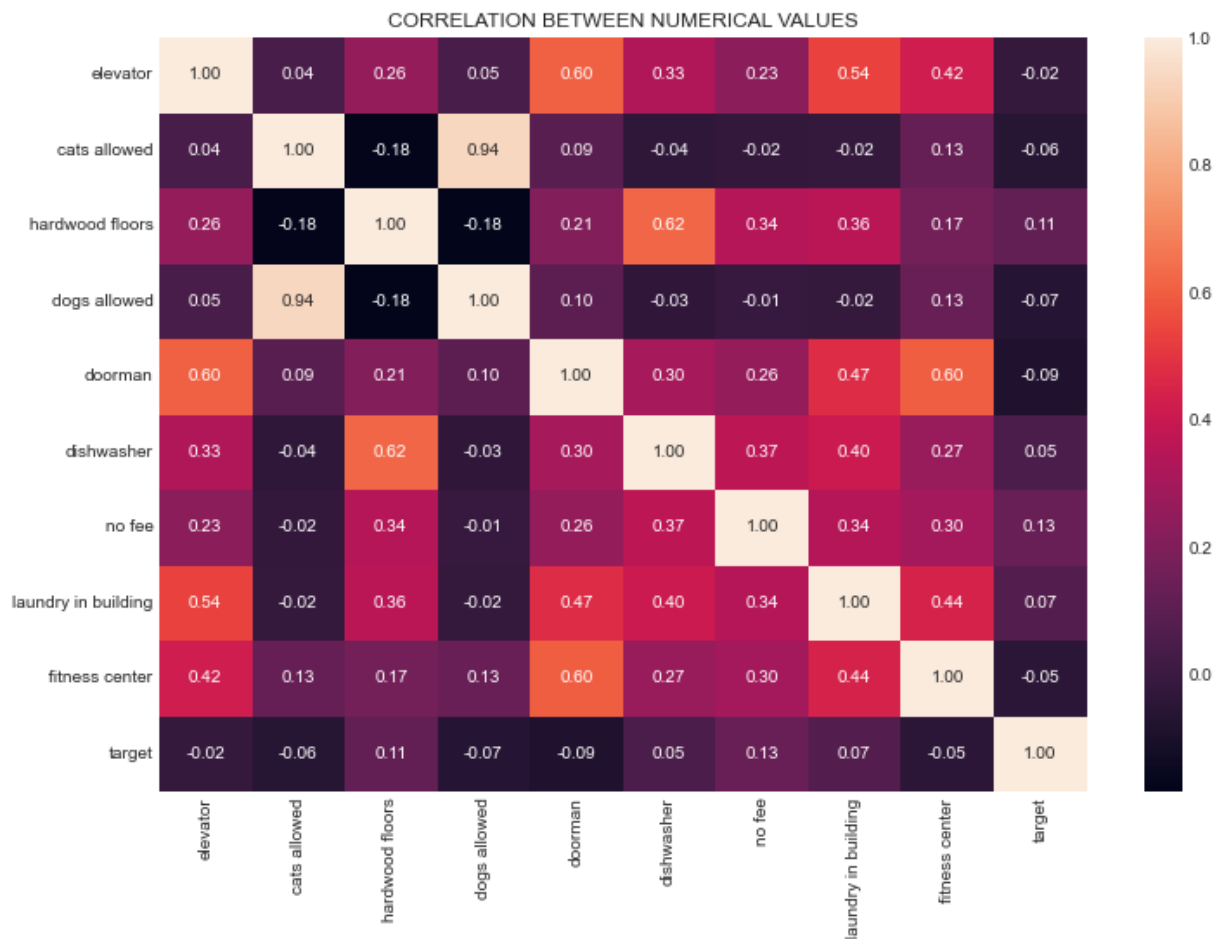
- As we can see low level is highly more than other interest levels

In [15]:

```
plt.style.use("seaborn-whitegrid")
plt.figure(figsize=(12,8))
plt.title("CORRELATION BETWEEN NUMERICAL VALUES")
num_col = ["rooms", "num_photos", "num_features", "has_email", "has_phone", "price",
sns.heatmap(train_df[num_col].corr(), annot = True, fmt = ".2f")
plt.show()
```



```
In [16]: plt.style.use("seaborn-whitegrid")
plt.figure(figsize=(12,8))
plt.title("CORRELATION BETWEEN NUMERICAL VALUES")
num_col = ['elevator', 'cats allowed', 'hardwood floors', 'dogs allowed', 'doorman',
sns.heatmap(train_df[num_col].corr(), annot = True, fmt = ".2f")
plt.show()
```



In [17]:

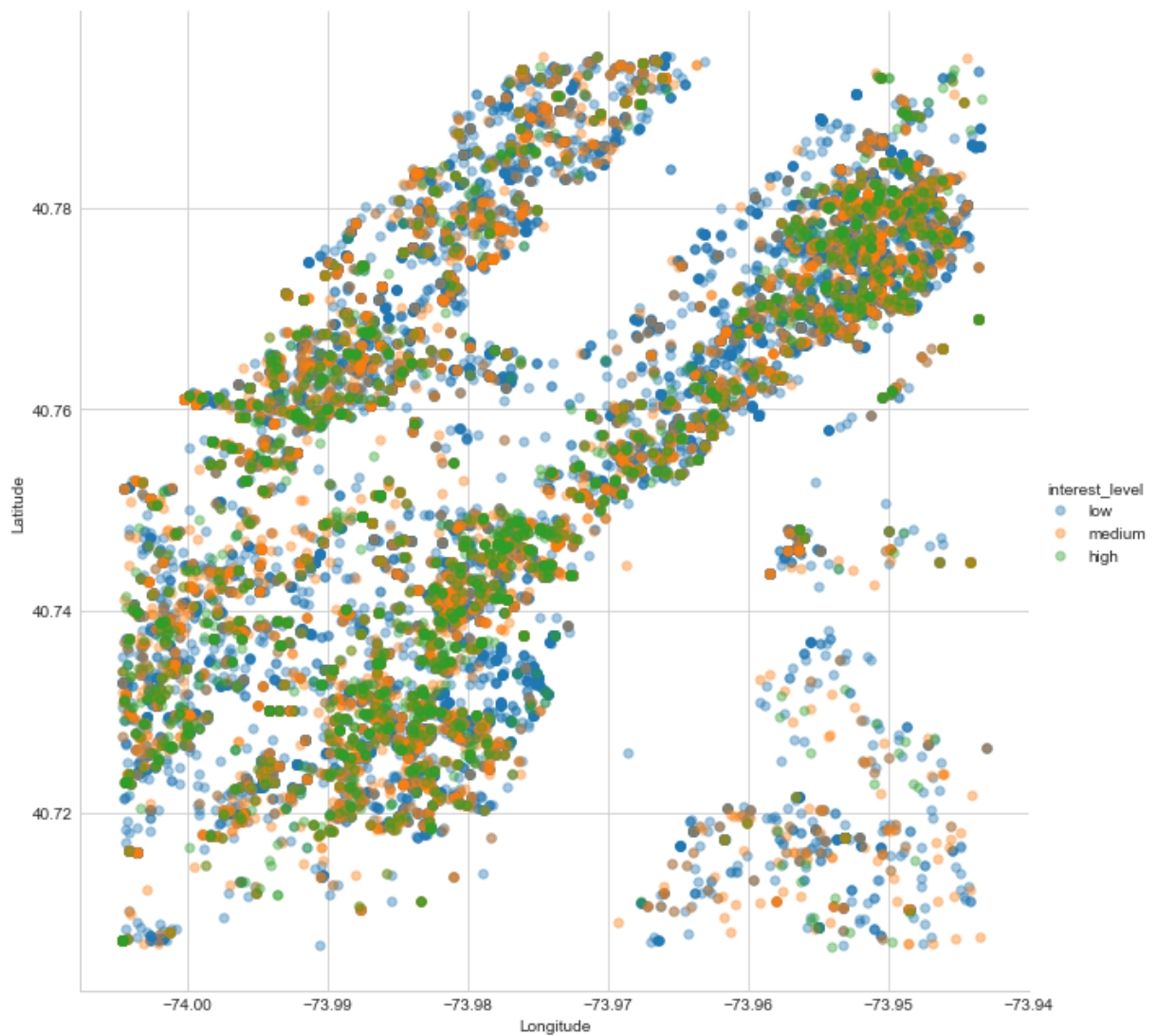
```

### Rent interest graph of New-York
sns.lmplot(x="longitude", y="latitude", fit_reg=False, hue='interest_level',
           hue_order=['low', 'medium', 'high'], size=9, scatter_kws={'alpha':0.4,'s'
           data=train_df[(train_df.longitude>train_df.longitude.quantile(0.1))
                           &(train_df.longitude<train_df.longitude.quantile(0.9))
                           &(train_df.latitude>train_df.latitude.quantile(0.1))
                           &(train_df.latitude<train_df.latitude.quantile(0.9))]);
plt.xlabel('Longitude');
plt.ylabel('Latitude');

```

C:\Users\burak\AppData\Roaming\Python\Python39\site-packages\seaborn\regression.py:581: UserWarning: The `size` parameter has been renamed to `height`; please update your code.

```
warnings.warn(msg, UserWarning)
```



In [18]:

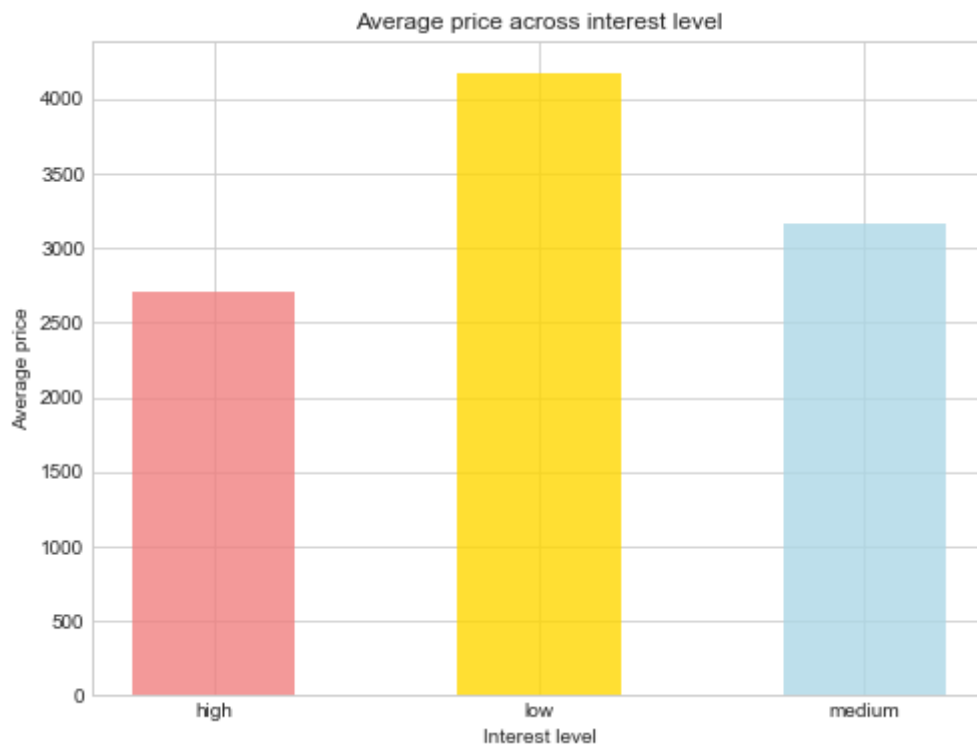
```

### Price exploration
prices=train_df.groupby('interest_level', as_index=False)['price'].mean()
colors = ['lightcoral', 'gold', 'lightblue']

fig=plt.figure(figsize=(8,6))
plt.bar(prices.interest_level, prices.price, color=colors, width=0.5, alpha=0.8)
#set titles
plt.xlabel('Interest level')
plt.ylabel('Average price')
plt.title('Average price across interest level')
plt.show()

```



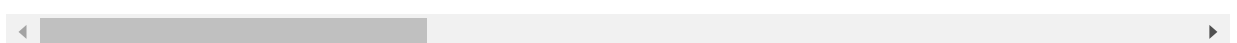


```
In [19]: train_df.groupby(['building_id', 'manager_id', 'interest_level']).count()
```

Out[19]:

				bathrooms
building_id		manager_id	interest_level	
0	001ce808ce1720e24a9510e014c69707		low	10
	003fc4e9a70053082f131b1054966aaf		low	1
	00a8d77892cab18fffaa22a751f1f8eb		low	3
	00f526d80353a50a52bbc26919e7ed5a		low	1
	01287194f20de51872e81f660def4784		low	8
...		...	...	...
ffde4f77049b631ee532fa9e0ebdd95d	41735645e0f8f13993c42894023f8e58		low	1
	f01d80f465348a6054bd7a9004af53b4		medium	1
ffe55387cd931c117ee1b8446f21953b	1a4dd24b22749ffaa3398b1b1b61c9ff		low	1
fff0a02ad82421c226c6d6765a0dde57	7efb139d6df9fc594c4db6ca96b8a1e7		low	1
fff5915444b98b72a44a9456901f083c	b944623e2af9b605eb0cba5236ee5f8e		low	2

33192 rows × 28 columns



- WORDCLOUD SHOWS US MOST FREQUENT WORDS IN THE DATASET, DEPENDS ON THE FREQUENCY WORDS SIZE IS GETTING BIGGER

```
In [20]: #WORDCLOUD FOR DESCRIPTION AND DISPLAY ADDRESS
#Preprocessing
text = ''
```

```
text_da = ''
text_desc = ''
text_str = ''
for ind, row in train_df.iterrows():
    for feature in row['features']:
        text = " ".join([text, " ".join(feature.strip().split(" "))])
    text_da = " ".join([text_da, " ".join(row['display_address'].strip().split(" "))])
    text_desc = " ".join([text_desc, row['description']])
    text_str = " ".join([text_str, row['street_address']])
text = text.strip()
text_da = text_da.strip()
text_desc = text_desc.strip()
text_str = text_str.strip()

# wordcloud for features
plt.figure(figsize=(12,6))
wordcloud = WordCloud(background_color='white', width=600, height=300, max_font_size
wordcloud.recolor(random_state=0)
plt.imshow(wordcloud)
plt.title("Wordcloud for features", fontsize=30)
plt.axis("off")
plt.show()

# wordcloud for display address
plt.figure(figsize=(12,6))
wordcloud = WordCloud(background_color='white', width=600, height=300, max_font_size
wordcloud.recolor(random_state=0)
plt.imshow(wordcloud)
plt.title("Wordcloud for Display Address", fontsize=30)
plt.axis("off")
plt.show()

# wordcloud for description
plt.figure(figsize=(12,6))
wordcloud = WordCloud(background_color='white', width=600, height=300, max_font_size
wordcloud.recolor(random_state=0)
plt.imshow(wordcloud)
plt.title("Wordcloud for Description", fontsize=30)
plt.axis("off")
plt.show()

# wordcloud for street address
plt.figure(figsize=(12,6))
wordcloud = WordCloud(background_color='white', width=600, height=300, max_font_size
wordcloud.recolor(random_state=0)
plt.imshow(wordcloud)
plt.title("Wordcloud for Street Address", fontsize=30)
plt.axis("off")
plt.show()
```

## Wordcloud for features

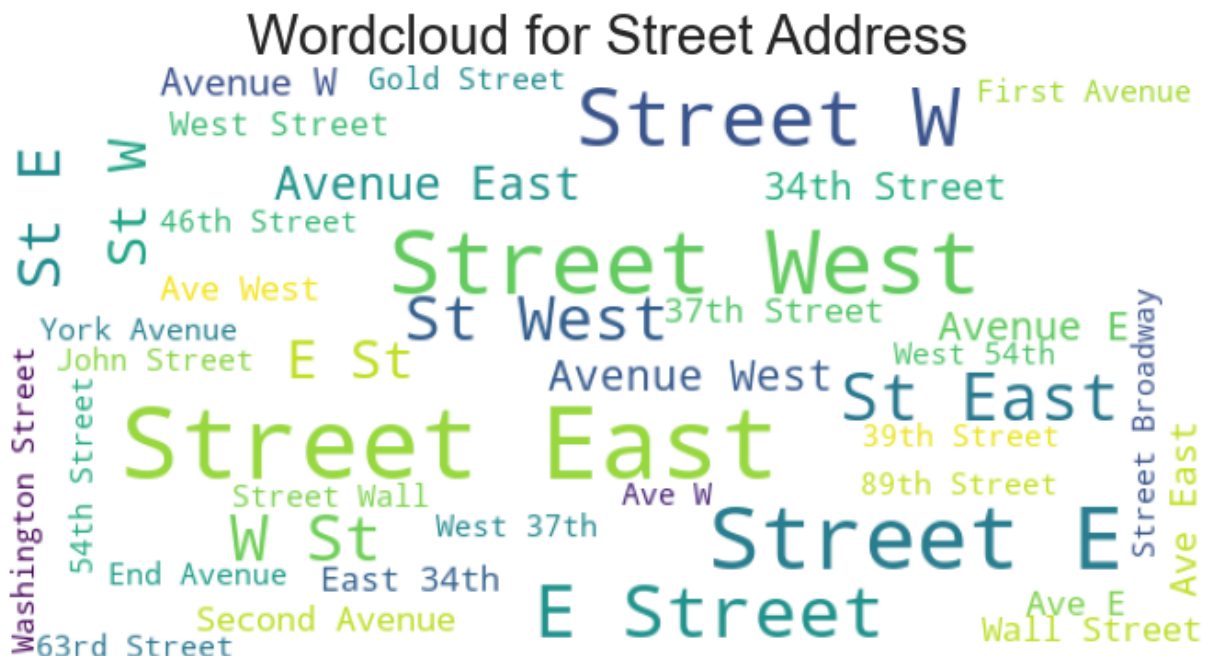


## Wordcloud for Display Address



## Wordcloud for Description





## DROP UNNECESSARY COLUMNS

```
In [21]: # TRAINING DATASET
train_df.drop('interest_level', axis=1, inplace=True)
train_df.drop('created', axis=1, inplace=True)
train_df.drop('description', axis=1, inplace=True)
train_df.drop('features', axis=1, inplace=True)
train_df.drop('photos', axis=1, inplace=True)

# TEST DATASET
test_df.drop('created', axis=1, inplace=True)
test_df.drop('description', axis=1, inplace=True)
test_df.drop('features', axis=1, inplace=True)
test_df.drop('photos', axis=1, inplace=True)
```

## LABEL ENCODING FOR CATEGORICAL VARIABLES

```
In [22]: categorical = ["display_address", "manager_id", "building_id", "street_address"]
for f in categorical:
    if train_df[f].dtype=='object':
        lbl = preprocessing.LabelEncoder()
        lbl.fit(list(train_df[f].values) + list(test_df[f].values))
        train_df[f] = lbl.transform(list(train_df[f].values))
        test_df[f] = lbl.transform(list(test_df[f].values))
```

## XGBOOST

```
In [23]: X = train_df.drop(['target'], axis = 1)
y = train_df.target
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size = .3,
                                                    random_state = 5,
                                                    stratify = y)

dtrain = xgb.DMatrix(X_train, label=y_train)
dvalid = xgb.DMatrix(X_test, label=y_test)
```

```
In [24]: kf = KFold(n_splits=5, shuffle=False)

X_train = X_train.values
y_train = y_train.values
scores = []

for train, test in kf.split(X_train, y_train):
    model = XGBClassifier(n_estimators=1000, learning_rate=0.05, max_depth = 10)
    model.fit(X_train[train], y_train[train])
    scores.append(model.score(X_train[test], y_train[test]))
```

C:\Users\burak\AppData\Roaming\Python\Python39\site-packages\xgboost\sklearn.py:114:6: UserWarning: The use of label encoder in XGBClassifier is deprecated and will be removed in a future release. To remove this warning, do the following: 1) Pass option use\_label\_encoder=False when constructing XGBClassifier object; and 2) Encode your labels (y) as integers starting with 0, i.e. 0, 1, 2, ..., [num\_class - 1].

warnings.warn(label\_encoder\_deprecation\_msg, UserWarning)  
[19:54:03] WARNING: C:/Users/Administrator/workspace/xgboost-win64\_release\_1.4.0/src/learner.cc:1095: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval\_metric if you'd like to restore the old behavior.

C:\Users\burak\AppData\Roaming\Python\Python39\site-packages\xgboost\sklearn.py:114:6: UserWarning: The use of label encoder in XGBClassifier is deprecated and will be removed in a future release. To remove this warning, do the following: 1) Pass option use\_label\_encoder=False when constructing XGBClassifier object; and 2) Encode your labels (y) as integers starting with 0, i.e. 0, 1, 2, ..., [num\_class - 1].

warnings.warn(label\_encoder\_deprecation\_msg, UserWarning)  
[19:55:56] WARNING: C:/Users/Administrator/workspace/xgboost-win64\_release\_1.4.0/src/learner.cc:1095: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval\_metric if you'd like to restore the old behavior.

C:\Users\burak\AppData\Roaming\Python\Python39\site-packages\xgboost\sklearn.py:114:6: UserWarning: The use of label encoder in XGBClassifier is deprecated and will be removed in a future release. To remove this warning, do the following: 1) Pass option use\_label\_encoder=False when constructing XGBClassifier object; and 2) Encode your labels (y) as integers starting with 0, i.e. 0, 1, 2, ..., [num\_class - 1].

warnings.warn(label\_encoder\_deprecation\_msg, UserWarning)  
[19:57:50] WARNING: C:/Users/Administrator/workspace/xgboost-win64\_release\_1.4.0/src/learner.cc:1095: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval\_metric if you'd like to restore the old behavior.

C:\Users\burak\AppData\Roaming\Python\Python39\site-packages\xgboost\sklearn.py:114:6: UserWarning: The use of label encoder in XGBClassifier is deprecated and will be removed in a future release. To remove this warning, do the following: 1) Pass option use\_label\_encoder=False when constructing XGBClassifier object; and 2) Encode your labels (y) as integers starting with 0, i.e. 0, 1, 2, ..., [num\_class - 1].

warnings.warn(label\_encoder\_deprecation\_msg, UserWarning)  
[19:59:41] WARNING: C:/Users/Administrator/workspace/xgboost-win64\_release\_1.4.0/src/learner.cc:1095: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval\_metric if you'd like to restore the old behavior.

C:\Users\burak\AppData\Roaming\Python\Python39\site-packages\xgboost\sklearn.py:114:6: UserWarning: The use of label encoder in XGBClassifier is deprecated and will be removed in a future release. To remove this warning, do the following: 1) Pass option use\_label\_encoder=False when constructing XGBClassifier object; and 2) Encode your labels (y) as integers starting with 0, i.e. 0, 1, 2, ..., [num\_class - 1].

warnings.warn(label\_encoder\_deprecation\_msg, UserWarning)  
[20:01:31] WARNING: C:/Users/Administrator/workspace/xgboost-win64\_release\_1.4.0/src/learner.cc:1095: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval\_metric if you'd like to restore the old behavior.

```
In [25]: def objective(trial):
        params = {
            'booster':trial.suggest_categorical('booster', ['gbtree', 'dart', 'gblinear',
            'learning_rate':trial.suggest_loguniform("learning_rate", 0.01, 0.1),
```



```

    'max_depth': trial.suggest_int("max_depth", 3, 11),
    'subsample': trial.suggest_uniform("subsample", 0.0, 1.0),
    'colsample_bytree': trial.suggest_uniform("colsample_bytree", 0.0, 1.0),
}

model = XGBClassifier(**params)
cv = KFold(n_splits=3, shuffle=True, random_state=None)
scorer = make_scorer(f1_score, greater_is_better=True)

bst = xgb.train(params, dtrain)
preds = bst.predict(dvalid)
pred_labels = np.rint(preds)
f1_scores = f1_score(y_test, pred_labels, average='micro')
return f1_scores

```

In [26]:

```

study = optuna.create_study(direction="maximize")
study.optimize(objective, n_trials=100, timeout=600)

```

[I 2021-09-28 20:03:24,539] A new study created in memory with name: no-name-daebf5a-f-b96e-4a59-b6ed-902aa132ab62

[I 2021-09-28 20:03:24,604] Trial 0 finished with value: 0.6946508172362555 and parameters: {'booster': 'gblinear', 'learning\_rate': 0.020292260641180088, 'max\_depth': 11, 'subsample': 0.9646159593651782, 'colsample\_bytree': 0.26122479208661264}. Best is trial 0 with value: 0.6946508172362555.

[I 2021-09-28 20:03:24,657] Trial 1 finished with value: 0.6948534377954884 and parameters: {'booster': 'gblinear', 'learning\_rate': 0.03971064264905856, 'max\_depth': 6, 'subsample': 0.9427153420848906, 'colsample\_bytree': 0.1649452021398924}. Best is trial 1 with value: 0.6948534377954884.

[20:03:24] WARNING: C:/Users/Administrator/workspace/xgboost-win64\_release\_1.4.0/src/learner.cc:573:

Parameters: { "colsample\_bytree", "max\_depth", "subsample" } might not be used.

This may not be accurate due to some parameters are only used in language bindings but

passed down to XGBoost core. Or some parameters are not used but slip through this

verification. Please open an issue if you find above cases.

[20:03:24] WARNING: C:/Users/Administrator/workspace/xgboost-win64\_release\_1.4.0/src/learner.cc:573:

Parameters: { "colsample\_bytree", "max\_depth", "subsample" } might not be used.

This may not be accurate due to some parameters are only used in language bindings but

passed down to XGBoost core. Or some parameters are not used but slip through this

verification. Please open an issue if you find above cases.

[I 2021-09-28 20:03:25,033] Trial 2 finished with value: 0.7147102526002973 and parameters: {'booster': 'gbtree', 'learning\_rate': 0.0671745658414657, 'max\_depth': 8, 'subsample': 0.6354989062342745, 'colsample\_bytree': 0.6553132346094781}. Best is trial 2 with value: 0.7147102526002973.

[I 2021-09-28 20:03:25,450] Trial 3 finished with value: 0.7099148993651221 and parameters: {'booster': 'dart', 'learning\_rate': 0.0169237910821856, 'max\_depth': 10, 'subsample': 0.4069375142820816, 'colsample\_bytree': 0.32822587227993383}. Best is trial 2 with value: 0.7147102526002973.

[I 2021-09-28 20:03:25,502] Trial 4 finished with value: 0.687221396731055 and parameters: {'booster': 'gblinear', 'learning\_rate': 0.08821518846399543, 'max\_depth': 3, 'subsample': 0.2375734164074117, 'colsample\_bytree': 0.9881174499320622}. Best is trial 2 with value: 0.7147102526002973.

[20:03:25] WARNING: C:/Users/Administrator/workspace/xgboost-win64\_release\_1.4.0/src/learner.cc:573:

Parameters: { "colsample\_bytree", "max\_depth", "subsample" } might not be used.

This may not be accurate due to some parameters are only used in language bindings but passed down to XGBoost core. Or some parameters are not used but slip through this verification. Please open an issue if you find above cases.

[I 2021-09-28 20:03:26,027] Trial 5 finished with value: 0.7138997703633662 and parameters: {'booster': 'gbtree', 'learning\_rate': 0.02385104822426317, 'max\_depth': 10, 'subsample': 0.7450124631332764, 'colsample\_bytree': 0.8659665505600008}. Best is trial 2 with value: 0.7147102526002973.

[I 2021-09-28 20:03:26,144] Trial 6 finished with value: 0.6794542752937998 and parameters: {'booster': 'gbtree', 'learning\_rate': 0.08883641577400002, 'max\_depth': 6, 'subsample': 0.05337516976559675, 'colsample\_bytree': 0.09281149045173487}. Best is trial 2 with value: 0.7147102526002973.

[I 2021-09-28 20:03:26,295] Trial 7 finished with value: 0.6928947723895718 and parameters: {'booster': 'gbtree', 'learning\_rate': 0.013574350314016949, 'max\_depth': 3, 'subsample': 0.42068000662224314, 'colsample\_bytree': 0.6816829496135494}. Best is trial 2 with value: 0.7147102526002973.

[I 2021-09-28 20:03:26,546] Trial 8 finished with value: 0.7041064433337836 and parameters: {'booster': 'dart', 'learning\_rate': 0.03234096957832758, 'max\_depth': 5, 'subsample': 0.9460704662766006, 'colsample\_bytree': 0.33868535156428325}. Best is trial 2 with value: 0.7147102526002973.

[I 2021-09-28 20:03:26,597] Trial 9 finished with value: 0.684857490206673 and parameters: {'booster': 'gblinear', 'learning\_rate': 0.09262038725003763, 'max\_depth': 11, 'subsample': 0.43073960277127077, 'colsample\_bytree': 0.868453982191589}. Best is trial 2 with value: 0.7147102526002973.

[20:03:26] WARNING: C:/Users/Administrator/workspace/xgboost-win64\_release\_1.4.0/src/learner.cc:573:

Parameters: { "colsample\_bytree", "max\_depth", "subsample" } might not be used.

This may not be accurate due to some parameters are only used in language bindings but passed down to XGBoost core. Or some parameters are not used but slip through this verification. Please open an issue if you find above cases.

[I 2021-09-28 20:03:26,963] Trial 10 finished with value: 0.7116709442118059 and parameters: {'booster': 'gbtree', 'learning\_rate': 0.050320430923254944, 'max\_depth': 8, 'subsample': 0.645510683957235, 'colsample\_bytree': 0.5290571385336701}. Best is trial 2 with value: 0.7147102526002973.

[I 2021-09-28 20:03:27,429] Trial 11 finished with value: 0.7138997703633662 and parameters: {'booster': 'gbtree', 'learning\_rate': 0.022554773252792207, 'max\_depth': 9, 'subsample': 0.6911471222635495, 'colsample\_bytree': 0.6931318662683087}. Best is trial 2 with value: 0.7147102526002973.

[I 2021-09-28 20:03:27,822] Trial 12 finished with value: 0.712481426448737 and parameters: {'booster': 'gbtree', 'learning\_rate': 0.05274892043099996, 'max\_depth': 8, 'subsample': 0.6809253650884115, 'colsample\_bytree': 0.7610409942075269}. Best is trial 2 with value: 0.7147102526002973.

[I 2021-09-28 20:03:28,270] Trial 13 finished with value: 0.7137646899905444 and parameters: {'booster': 'gbtree', 'learning\_rate': 0.02662884254569593, 'max\_depth': 9, 'subsample': 0.7837224447296913, 'colsample\_bytree': 0.5626004791387937}. Best is trial 2 with value: 0.7147102526002973.

[I 2021-09-28 20:03:28,731] Trial 14 finished with value: 0.7083614750776712 and parameters: {'booster': 'gbtree', 'learning\_rate': 0.011059766201976469, 'max\_depth': 9, 'subsample': 0.5665889635994922, 'colsample\_bytree': 0.9887478699978333}. Best is trial 2 with value: 0.7147102526002973.

[I 2021-09-28 20:03:29,085] Trial 15 finished with value: 0.710928002161286 and parameters: {'booster': 'gbtree', 'learning\_rate': 0.05756558771096872, 'max\_depth': 7, 'subsample': 0.7911312284366812, 'colsample\_bytree': 0.8100122440319104}. Best is trial 2 with value: 0.7147102526002973.

[I 2021-09-28 20:03:29,649] Trial 16 finished with value: 0.7191679049034175 and parameters: {'booster': 'dart', 'learning\_rate': 0.035264684781289274, 'max\_depth': 10, 'subsample': 0.8067133889029242, 'colsample\_bytree': 0.6122244968693766}. Best is trial 16 with value: 0.7191679049034175.

[I 2021-09-28 20:03:30,051] Trial 17 finished with value: 0.7128866675672024 and parameters: {'booster': 'dart', 'learning\_rate': 0.06836556532618639, 'max\_depth': 8, 'subsample': 0.8486250199709774, 'colsample\_bytree': 0.4473403283188177}. Best is trial 2 with value: 0.7147102526002973.

rial 16 with value: 0.7191679049034175.

[I 2021-09-28 20:03:30,580] Trial 18 finished with value: 0.7159935161421046 and parameters: {'booster': 'dart', 'learning\_rate': 0.037320689194929785, 'max\_depth': 10, 'subsample': 0.5638404349998636, 'colsample\_bytree': 0.6215079996642638}. Best is trial 16 with value: 0.7191679049034175.

[I 2021-09-28 20:03:30,964] Trial 19 finished with value: 0.7096447386194786 and parameters: {'booster': 'dart', 'learning\_rate': 0.0371396354343268, 'max\_depth': 10, 'subsample': 0.23224921685508604, 'colsample\_bytree': 0.4591261929617707}. Best is trial 16 with value: 0.7191679049034175.

[I 2021-09-28 20:03:31,535] Trial 20 finished with value: 0.7143050114818317 and parameters: {'booster': 'dart', 'learning\_rate': 0.04155271772780083, 'max\_depth': 11, 'subsample': 0.5348623880874096, 'colsample\_bytree': 0.5590648212611942}. Best is trial 16 with value: 0.7191679049034175.

[I 2021-09-28 20:03:32,062] Trial 21 finished with value: 0.7159935161421046 and parameters: {'booster': 'dart', 'learning\_rate': 0.06627777083389942, 'max\_depth': 10, 'subsample': 0.5896535935142201, 'colsample\_bytree': 0.6360576455659642}. Best is trial 16 with value: 0.7191679049034175.

[I 2021-09-28 20:03:32,553] Trial 22 finished with value: 0.7097122788058896 and parameters: {'booster': 'dart', 'learning\_rate': 0.031074603153888544, 'max\_depth': 10, 'subsample': 0.29207380364473673, 'colsample\_bytree': 0.6266695236990383}. Best is trial 16 with value: 0.7191679049034175.

[I 2021-09-28 20:03:33,042] Trial 23 finished with value: 0.7136296096177226 and parameters: {'booster': 'dart', 'learning\_rate': 0.045415529004500334, 'max\_depth': 9, 'subsample': 0.5886950833271988, 'colsample\_bytree': 0.7503705986866542}. Best is trial 16 with value: 0.7191679049034175.

[I 2021-09-28 20:03:33,634] Trial 24 finished with value: 0.7130892881264352 and parameters: {'booster': 'dart', 'learning\_rate': 0.06807393973985171, 'max\_depth': 11, 'subsample': 0.8616114449530948, 'colsample\_bytree': 0.43556067572408863}. Best is trial 16 with value: 0.7191679049034175.

[I 2021-09-28 20:03:34,146] Trial 25 finished with value: 0.7145076320410645 and parameters: {'booster': 'dart', 'learning\_rate': 0.03328169118527605, 'max\_depth': 10, 'subsample': 0.4832966217444564, 'colsample\_bytree': 0.5891808135221773}. Best is trial 16 with value: 0.7191679049034175.

[I 2021-09-28 20:03:34,431] Trial 26 finished with value: 0.7062001891125219 and parameters: {'booster': 'dart', 'learning\_rate': 0.02709541586655903, 'max\_depth': 7, 'subsample': 0.3367935023455707, 'colsample\_bytree': 0.39385831992493436}. Best is trial 16 with value: 0.7191679049034175.

[I 2021-09-28 20:03:34,860] Trial 27 finished with value: 0.7125489666351479 and parameters: {'booster': 'dart', 'learning\_rate': 0.05988756308486751, 'max\_depth': 9, 'subsample': 0.5055132308717751, 'colsample\_bytree': 0.5055379249962058}. Best is trial 16 with value: 0.7191679049034175.

[I 2021-09-28 20:03:35,560] Trial 28 finished with value: 0.7182898824800756 and parameters: {'booster': 'dart', 'learning\_rate': 0.017133566176005623, 'max\_depth': 11, 'subsample': 0.8485225692561953, 'colsample\_bytree': 0.7226461493981079}. Best is trial 16 with value: 0.7191679049034175.

[I 2021-09-28 20:03:36,295] Trial 29 finished with value: 0.721194110495745 and parameters: {'booster': 'dart', 'learning\_rate': 0.018811415804001412, 'max\_depth': 11, 'subsample': 0.8815985967560147, 'colsample\_bytree': 0.749042232320869}. Best is trial 29 with value: 0.721194110495745.

[I 2021-09-28 20:03:36,377] Trial 30 finished with value: 0.6946508172362555 and parameters: {'booster': 'gblinear', 'learning\_rate': 0.016474307634597018, 'max\_depth': 11, 'subsample': 0.9958486279158816, 'colsample\_bytree': 0.8963573038976369}. Best is trial 29 with value: 0.721194110495745.

[20:03:36] WARNING: C:/Users/Administrator/workspace/xgboost-win64\_release\_1.4.0/src/learner.cc:573:

Parameters: { "colsample\_bytree", "max\_depth", "subsample" } might not be used.

This may not be accurate due to some parameters are only used in language bindings but

passed down to XGBoost core. Or some parameters are not used but slip through this

verification. Please open an issue if you find above cases.

[I 2021-09-28 20:03:37,122] Trial 31 finished with value: 0.718019721734432 and parameters: {'booster': 'dart', 'learning\_rate': 0.0181436050865597, 'max\_depth': 11, 'subsample': 0.8906079260823732, 'colsample\_bytree': 0.7384794446675637}. Best is trial 29 with value: 0.721194110495745.

[I 2021-09-28 20:03:37,816] Trial 32 finished with value: 0.7179521815480211 and par



ameters: {'booster': 'dart', 'learning\_rate': 0.01882680340509867, 'max\_depth': 11, 'subsample': 0.8661220162726049, 'colsample\_bytree': 0.7397357999905892}. Best is trial 29 with value: 0.721194110495745.

[I 2021-09-28 20:03:38,529] Trial 33 finished with value: 0.7192354450898284 and parameters: {'booster': 'dart', 'learning\_rate': 0.013538808018023997, 'max\_depth': 11, 'subsample': 0.9096978750081544, 'colsample\_bytree': 0.826835748868562}. Best is trial 29 with value: 0.721194110495745.

[I 2021-09-28 20:03:39,239] Trial 34 finished with value: 0.7201810076995813 and parameters: {'booster': 'dart', 'learning\_rate': 0.014230757875599928, 'max\_depth': 11, 'subsample': 0.9544959764128038, 'colsample\_bytree': 0.8147672941226156}. Best is trial 29 with value: 0.721194110495745.

[I 2021-09-28 20:03:39,988] Trial 35 finished with value: 0.713224368499257 and parameters: {'booster': 'dart', 'learning\_rate': 0.013822650045790318, 'max\_depth': 11, 'subsample': 0.9290873154704613, 'colsample\_bytree': 0.9371405712140983}. Best is trial 29 with value: 0.721194110495745.

[I 2021-09-28 20:03:40,062] Trial 36 finished with value: 0.6946508172362555 and parameters: {'booster': 'gblinear', 'learning\_rate': 0.010067746919951194, 'max\_depth': 10, 'subsample': 0.9974022892826917, 'colsample\_bytree': 0.8000956849039891}. Best is trial 29 with value: 0.721194110495745.

[20:03:40] WARNING: C:/Users/Administrator/workspace/xgboost-win64\_release\_1.4.0/src/learner.cc:573:

Parameters: { "colsample\_bytree", "max\_depth", "subsample" } might not be used.

This may not be accurate due to some parameters are only used in language bindings but

passed down to XGBoost core. Or some parameters are not used but slip through this

verification. Please open an issue if you find above cases.

[I 2021-09-28 20:03:40,793] Trial 37 finished with value: 0.7169390787518574 and parameters: {'booster': 'dart', 'learning\_rate': 0.01355911140997535, 'max\_depth': 11, 'subsample': 0.7771241734386887, 'colsample\_bytree': 0.8300727501671011}. Best is trial 29 with value: 0.721194110495745.

[I 2021-09-28 20:03:41,113] Trial 38 finished with value: 0.6970147237606376 and parameters: {'booster': 'dart', 'learning\_rate': 0.011906422502053116, 'max\_depth': 4, 'subsample': 0.9274409070115799, 'colsample\_bytree': 0.9430271432831436}. Best is trial 29 with value: 0.721194110495745.

[I 2021-09-28 20:03:41,193] Trial 39 finished with value: 0.6946508172362555 and parameters: {'booster': 'gblinear', 'learning\_rate': 0.02142601599981189, 'max\_depth': 10, 'subsample': 0.7521829934790993, 'colsample\_bytree': 0.7950583896644295}. Best is trial 29 with value: 0.721194110495745.

[20:03:41] WARNING: C:/Users/Administrator/workspace/xgboost-win64\_release\_1.4.0/src/learner.cc:573:

Parameters: { "colsample\_bytree", "max\_depth", "subsample" } might not be used.

This may not be accurate due to some parameters are only used in language bindings but

passed down to XGBoost core. Or some parameters are not used but slip through this

verification. Please open an issue if you find above cases.

[I 2021-09-28 20:03:41,407] Trial 40 finished with value: 0.6930298527623936 and parameters: {'booster': 'dart', 'learning\_rate': 0.014738804701204717, 'max\_depth': 6, 'subsample': 0.8181107078519407, 'colsample\_bytree': 0.01588622442765697}. Best is trial 29 with value: 0.721194110495745.

[I 2021-09-28 20:03:42,139] Trial 41 finished with value: 0.7176820208023774 and parameters: {'booster': 'dart', 'learning\_rate': 0.016298016602411536, 'max\_depth': 11, 'subsample': 0.9181864988252767, 'colsample\_bytree': 0.6777726503532263}. Best is trial 29 with value: 0.721194110495745.

[I 2021-09-28 20:03:42,858] Trial 42 finished with value: 0.7161961367013372 and parameters: {'booster': 'dart', 'learning\_rate': 0.012283881973886815, 'max\_depth': 11, 'subsample': 0.8376319583898734, 'colsample\_bytree': 0.6991834958272862}. Best is trial 29 with value: 0.721194110495745.

[I 2021-09-28 20:03:43,519] Trial 43 finished with value: 0.7131568283128461 and parameters: {'booster': 'dart', 'learning\_rate': 0.0151500778028335, 'max\_depth': 10, 'subsample': 0.9056849499132963, 'colsample\_bytree': 0.8771653449703377}. Best is trial 29 with value: 0.721194110495745.

```
[I 2021-09-28 20:03:44,206] Trial 44 finished with value: 0.718897744157774 and parameters: {'booster': 'dart', 'learning_rate': 0.02005176020136941, 'max_depth': 11, 'subsample': 0.705346231403215, 'colsample_bytree': 0.7252472854272886}. Best is trial 29 with value: 0.721194110495745.
[I 2021-09-28 20:03:44,852] Trial 45 finished with value: 0.7163312170741591 and parameters: {'booster': 'dart', 'learning_rate': 0.019256851108760055, 'max_depth': 10, 'subsample': 0.7342760049214084, 'colsample_bytree': 0.843112713386392}. Best is trial 29 with value: 0.721194110495745.
[I 2021-09-28 20:03:45,229] Trial 46 finished with value: 0.7103201404835877 and parameters: {'booster': 'dart', 'learning_rate': 0.02042733920615968, 'max_depth': 9, 'subsample': 0.6892426288076028, 'colsample_bytree': 0.22796477375413676}. Best is trial 29 with value: 0.721194110495745.
[I 2021-09-28 20:03:45,307] Trial 47 finished with value: 0.6946508172362555 and parameters: {'booster': 'gblinear', 'learning_rate': 0.025147714537105348, 'max_depth': 5, 'subsample': 0.0028050626524913636, 'colsample_bytree': 0.9190239792096999}. Best is trial 29 with value: 0.721194110495745.
[20:03:45] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.4.0/src/learner.cc:573:
Parameters: { "colsample_bytree", "max_depth", "subsample" } might not be used.
```

This may not be accurate due to some parameters are only used in language bindings but passed down to XGBoost core. Or some parameters are not used but slip through this verification. Please open an issue if you find above cases.

```
[I 2021-09-28 20:03:46,070] Trial 48 finished with value: 0.7216668918006215 and parameters: {'booster': 'dart', 'learning_rate': 0.012372918087737834, 'max_depth': 11, 'subsample': 0.9655743598779992, 'colsample_bytree': 0.7652990096612262}. Best is trial 48 with value: 0.7216668918006215.
[I 2021-09-28 20:03:46,734] Trial 49 finished with value: 0.7160610563285155 and parameters: {'booster': 'dart', 'learning_rate': 0.012303351988637181, 'max_depth': 10, 'subsample': 0.958911401277666, 'colsample_bytree': 0.7820791613601907}. Best is trial 48 with value: 0.7216668918006215.
[I 2021-09-28 20:03:47,455] Trial 50 finished with value: 0.7201810076995813 and parameters: {'booster': 'dart', 'learning_rate': 0.010151891782989768, 'max_depth': 11, 'subsample': 0.9649906452644258, 'colsample_bytree': 0.66182652743338}. Best is trial 48 with value: 0.7216668918006215.
[I 2021-09-28 20:03:48,178] Trial 51 finished with value: 0.717209239497501 and parameters: {'booster': 'dart', 'learning_rate': 0.011307070282379946, 'max_depth': 11, 'subsample': 0.9755655344482176, 'colsample_bytree': 0.6551215737647302}. Best is trial 48 with value: 0.7216668918006215.
[I 2021-09-28 20:03:48,905] Trial 52 finished with value: 0.7203160880724031 and parameters: {'booster': 'dart', 'learning_rate': 0.010474055695772665, 'max_depth': 11, 'subsample': 0.9542306953682429, 'colsample_bytree': 0.5924487961313701}. Best is trial 48 with value: 0.7216668918006215.
[I 2021-09-28 20:03:49,661] Trial 53 finished with value: 0.7189652843441848 and parameters: {'booster': 'dart', 'learning_rate': 0.010331062316638428, 'max_depth': 11, 'subsample': 0.959888569030375, 'colsample_bytree': 0.835803971246517}. Best is trial 48 with value: 0.7216668918006215.
[I 2021-09-28 20:03:50,368] Trial 54 finished with value: 0.7185600432257193 and parameters: {'booster': 'dart', 'learning_rate': 0.013021639884035334, 'max_depth': 11, 'subsample': 0.8896060723033562, 'colsample_bytree': 0.665962629167435}. Best is trial 48 with value: 0.7216668918006215.
[I 2021-09-28 20:03:50,944] Trial 55 finished with value: 0.7169390787518574 and parameters: {'booster': 'dart', 'learning_rate': 0.01084308516778214, 'max_depth': 10, 'subsample': 0.9530968202470905, 'colsample_bytree': 0.5494958310877055}. Best is trial 48 with value: 0.7216668918006215.
[I 2021-09-28 20:03:51,658] Trial 56 finished with value: 0.7186275834121301 and parameters: {'booster': 'gbtree', 'learning_rate': 0.015136326192996766, 'max_depth': 11, 'subsample': 0.8917466343004545, 'colsample_bytree': 0.7706153694063728}. Best is trial 48 with value: 0.7216668918006215.
[I 2021-09-28 20:03:52,312] Trial 57 finished with value: 0.718762663784952 and parameters: {'booster': 'dart', 'learning_rate': 0.011745983900707916, 'max_depth': 10, 'subsample': 0.9848787074282745, 'colsample_bytree': 0.7071927933069218}. Best is trial 48 with value: 0.7216668918006215.
[I 2021-09-28 20:03:52,635] Trial 58 finished with value: 0.7106578414156423 and parameters: {'booster': 'dart', 'learning_rate': 0.012846900135924792, 'max_depth': 9,
```

'subsample': 0.14330199478783023, 'colsample\_bytree': 0.5913497292620224}. Best is trial 48 with value: 0.7216668918006215.

[I 2021-09-28 20:03:53,406] Trial 59 finished with value: 0.7104552208564094 and parameters: {'booster': 'dart', 'learning\_rate': 0.01455725277492492, 'max\_depth': 11, 'subsample': 0.9314834471900003, 'colsample\_bytree': 0.9684848901674903}. Best is trial 48 with value: 0.7216668918006215.

[I 2021-09-28 20:03:54,102] Trial 60 finished with value: 0.717276779683912 and parameters: {'booster': 'gbtree', 'learning\_rate': 0.010678861708340573, 'max\_depth': 11, 'subsample': 0.8727493585619437, 'colsample\_bytree': 0.8764135601270095}. Best is trial 48 with value: 0.7216668918006215.

[I 2021-09-28 20:03:54,642] Trial 61 finished with value: 0.7197757665811158 and parameters: {'booster': 'dart', 'learning\_rate': 0.011286341663873205, 'max\_depth': 10, 'subsample': 0.830295990985018, 'colsample\_bytree': 0.596781608987933}. Best is trial 48 with value: 0.7216668918006215.

[I 2021-09-28 20:03:55,306] Trial 62 finished with value: 0.7183574226664866 and parameters: {'booster': 'dart', 'learning\_rate': 0.010024346158224765, 'max\_depth': 11, 'subsample': 0.8259038784521531, 'colsample\_bytree': 0.5922723455953475}. Best is trial 48 with value: 0.7216668918006215.

[I 2021-09-28 20:03:55,996] Trial 63 finished with value: 0.7183574226664866 and parameters: {'booster': 'dart', 'learning\_rate': 0.011426413043885426, 'max\_depth': 10, 'subsample': 0.9997344078224242, 'colsample\_bytree': 0.6559447552602073}. Best is trial 48 with value: 0.7216668918006215.

[I 2021-09-28 20:03:56,491] Trial 64 finished with value: 0.7134269890584898 and parameters: {'booster': 'dart', 'learning\_rate': 0.01308911973031348, 'max\_depth': 8, 'subsample': 0.9507203367235719, 'colsample\_bytree': 0.7688425548216874}. Best is trial 48 with value: 0.7216668918006215.

[I 2021-09-28 20:03:57,067] Trial 65 finished with value: 0.7170741591246792 and parameters: {'booster': 'dart', 'learning\_rate': 0.014156712721780911, 'max\_depth': 10, 'subsample': 0.7977218391884293, 'colsample\_bytree': 0.6916553328811582}. Best is trial 48 with value: 0.7216668918006215.

[I 2021-09-28 20:03:57,141] Trial 66 finished with value: 0.6946508172362555 and parameters: {'booster': 'gblinear', 'learning\_rate': 0.015879318345781386, 'max\_depth': 11, 'subsample': 0.9114135822002497, 'colsample\_bytree': 0.5314582032603137}. Best is trial 48 with value: 0.7216668918006215.

[20:03:57] WARNING: C:/Users/Administrator/workspace/xgboost-win64\_release\_1.4.0/src/learner.cc:573:  
Parameters: { "colsample\_bytree", "max\_depth", "subsample" } might not be used.

This may not be accurate due to some parameters are only used in language bindings but passed down to XGBoost core. Or some parameters are not used but slip through this verification. Please open an issue if you find above cases.

[I 2021-09-28 20:03:57,792] Trial 67 finished with value: 0.7184249628528975 and parameters: {'booster': 'dart', 'learning\_rate': 0.017791349917217533, 'max\_depth': 11, 'subsample': 0.8626376949829242, 'colsample\_bytree': 0.4828850535803974}. Best is trial 48 with value: 0.7216668918006215.

[I 2021-09-28 20:03:58,523] Trial 68 finished with value: 0.7189652843441848 and parameters: {'booster': 'dart', 'learning\_rate': 0.011306096410324385, 'max\_depth': 11, 'subsample': 0.6384818616599034, 'colsample\_bytree': 0.8173956793520236}. Best is trial 48 with value: 0.7216668918006215.

[I 2021-09-28 20:03:59,099] Trial 69 finished with value: 0.7180872619208428 and parameters: {'booster': 'dart', 'learning\_rate': 0.01260064570356716, 'max\_depth': 10, 'subsample': 0.9674580868914424, 'colsample\_bytree': 0.6229402938318169}. Best is trial 48 with value: 0.7216668918006215.

[I 2021-09-28 20:03:59,669] Trial 70 finished with value: 0.7157908955828718 and parameters: {'booster': 'dart', 'learning\_rate': 0.01070752712462521, 'max\_depth': 9, 'subsample': 0.893915085729417, 'colsample\_bytree': 0.8475726160246746}. Best is trial 48 with value: 0.7216668918006215.

[I 2021-09-28 20:04:00,248] Trial 71 finished with value: 0.7181548021072538 and parameters: {'booster': 'dart', 'learning\_rate': 0.030216664700848193, 'max\_depth': 10, 'subsample': 0.8063933463206375, 'colsample\_bytree': 0.6000755710662812}. Best is trial 48 with value: 0.7216668918006215.

[I 2021-09-28 20:04:00,856] Trial 72 finished with value: 0.7182223422936647 and parameters: {'booster': 'dart', 'learning\_rate': 0.040394626854625616, 'max\_depth': 11, 'subsample': 0.7633199889553639, 'colsample\_bytree': 0.5158080144850035}. Best is trial 48 with value: 0.7216668918006215.

```
[I 2021-09-28 20:04:01,460] Trial 73 finished with value: 0.7170066189382682 and par
ameters: {'booster': 'dart', 'learning_rate': 0.036368746862164764, 'max_depth': 10,
'subsample': 0.8305930068316546, 'colsample_bytree': 0.751084982638813}. Best is tr
ial 48 with value: 0.7216668918006215.
[I 2021-09-28 20:04:02,114] Trial 74 finished with value: 0.7178171011751993 and par
ameters: {'booster': 'dart', 'learning_rate': 0.02936970644564743, 'max_depth': 11,
'subsample': 0.9315417966738104, 'colsample_bytree': 0.5590706605419127}. Best is t
rial 48 with value: 0.7216668918006215.
[I 2021-09-28 20:04:02,781] Trial 75 finished with value: 0.7190328245305957 and par
ameters: {'booster': 'dart', 'learning_rate': 0.03430264128374902, 'max_depth': 11,
'subsample': 0.7227826736867355, 'colsample_bytree': 0.6474559518754931}. Best is t
rial 48 with value: 0.7216668918006215.
[I 2021-09-28 20:04:03,357] Trial 76 finished with value: 0.7157908955828718 and par
ameters: {'booster': 'gbtree', 'learning_rate': 0.011972025493350353, 'max_depth': 1
0, 'subsample': 0.8573812263236307, 'colsample_bytree': 0.720576827385284}. Best is
trial 48 with value: 0.7216668918006215.
[I 2021-09-28 20:04:03,813] Trial 77 finished with value: 0.7091719573146021 and par
ameters: {'booster': 'dart', 'learning_rate': 0.013938185498276776, 'max_depth': 7,
'subsample': 0.8795841450966531, 'colsample_bytree': 0.6072320984659302}. Best is t
rial 48 with value: 0.7216668918006215.
[I 2021-09-28 20:04:03,905] Trial 78 finished with value: 0.6961367013372957 and par
ameters: {'booster': 'gblinear', 'learning_rate': 0.04705477528882798, 'max_depth':
9, 'subsample': 0.9135763909649639, 'colsample_bytree': 0.4070810794721377}. Best i
s trial 48 with value: 0.7216668918006215.
[20:04:03] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.4.0/sr
c/learner.cc:573:
Parameters: { "colsample_bytree", "max_depth", "subsample" } might not be used.
```

This may not be accurate due to some parameters are only used in language bindings but passed down to XGBoost core. Or some parameters are not used but slip through this verification. Please open an issue if you find above cases.

```
[I 2021-09-28 20:04:04,776] Trial 79 finished with value: 0.7199783871403485 and par
ameters: {'booster': 'dart', 'learning_rate': 0.010913186886966556, 'max_depth': 11,
'subsample': 0.9765155569894821, 'colsample_bytree': 0.8008429505625304}. Best is tr
ial 48 with value: 0.7216668918006215.
[I 2021-09-28 20:04:05,565] Trial 80 finished with value: 0.718830203971363 and para
meters: {'booster': 'dart', 'learning_rate': 0.01557670217818206, 'max_depth': 11,
'subsample': 0.9784311359480626, 'colsample_bytree': 0.8016183517419672}. Best is t
rial 48 with value: 0.7216668918006215.
[I 2021-09-28 20:04:06,281] Trial 81 finished with value: 0.7174794002431447 and par
ameters: {'booster': 'dart', 'learning_rate': 0.010478721830830356, 'max_depth': 11,
'subsample': 0.9474722341024752, 'colsample_bytree': 0.6921893041484868}. Best is tr
ial 48 with value: 0.7216668918006215.
[I 2021-09-28 20:04:07,125] Trial 82 finished with value: 0.7174794002431447 and par
ameters: {'booster': 'dart', 'learning_rate': 0.013536126485901833, 'max_depth': 11,
'subsample': 0.9999000348910785, 'colsample_bytree': 0.8910071441332547}. Best is tr
ial 48 with value: 0.7216668918006215.
[I 2021-09-28 20:04:07,947] Trial 83 finished with value: 0.7201810076995813 and par
ameters: {'booster': 'dart', 'learning_rate': 0.011077130993983188, 'max_depth': 11,
'subsample': 0.927099820358151, 'colsample_bytree': 0.8557942716081728}. Best is tri
al 48 with value: 0.7216668918006215.
[I 2021-09-28 20:04:08,745] Trial 84 finished with value: 0.7194380656490612 and par
ameters: {'booster': 'dart', 'learning_rate': 0.011078087297627033, 'max_depth': 11,
'subsample': 0.9375296465879259, 'colsample_bytree': 0.7840254614407685}. Best is tr
ial 48 with value: 0.7216668918006215.
[I 2021-09-28 20:04:09,540] Trial 85 finished with value: 0.7197082263947049 and par
ameters: {'booster': 'dart', 'learning_rate': 0.011544035913789054, 'max_depth': 11,
'subsample': 0.9391583145316493, 'colsample_bytree': 0.8624582863442294}. Best is tr
ial 48 with value: 0.7216668918006215.
[I 2021-09-28 20:04:10,371] Trial 86 finished with value: 0.7176820208023774 and par
ameters: {'booster': 'dart', 'learning_rate': 0.010020077021367408, 'max_depth': 11,
'subsample': 0.9640730810674187, 'colsample_bytree': 0.9156502078807611}. Best is tr
ial 48 with value: 0.7216668918006215.
[I 2021-09-28 20:04:11,003] Trial 87 finished with value: 0.7166013778198028 and par
ameters: {'booster': 'dart', 'learning_rate': 0.012048605634647058, 'max_depth': 11,
```



'subsample': 0.36475416753586687, 'colsample\_bytree': 0.745863781107261}. Best is trial 48 with value: 0.7216668918006215.

[I 2021-09-28 20:04:11,766] Trial 88 finished with value: 0.7132919086856679 and parameters: {'booster': 'dart', 'learning\_rate': 0.011566489404233065, 'max\_depth': 10, 'subsample': 0.9072847228486199, 'colsample\_bytree': 0.8587412189235812}. Best is trial 48 with value: 0.7216668918006215.

[I 2021-09-28 20:04:12,721] Trial 89 finished with value: 0.7195056058354721 and parameters: {'booster': 'dart', 'learning\_rate': 0.012927004470023944, 'max\_depth': 11, 'subsample': 0.973012402166559, 'colsample\_bytree': 0.8187632794905692}. Best is trial 48 with value: 0.7216668918006215.

[I 2021-09-28 20:04:13,392] Trial 90 finished with value: 0.707888693772795 and parameters: {'booster': 'gbtree', 'learning\_rate': 0.010728910620002686, 'max\_depth': 10, 'subsample': 0.8549689908731146, 'colsample\_bytree': 0.9624793481703484}. Best is trial 48 with value: 0.7216668918006215.

[I 2021-09-28 20:04:14,190] Trial 91 finished with value: 0.7207213291908686 and parameters: {'booster': 'dart', 'learning\_rate': 0.012601114272611682, 'max\_depth': 11, 'subsample': 0.976876154702135, 'colsample\_bytree': 0.8158769381874732}. Best is trial 48 with value: 0.7216668918006215.

[I 2021-09-28 20:04:15,082] Trial 92 finished with value: 0.7174118600567337 and parameters: {'booster': 'dart', 'learning\_rate': 0.012294412984961811, 'max\_depth': 11, 'subsample': 0.9496196294640821, 'colsample\_bytree': 0.9000550873532362}. Best is trial 48 with value: 0.7216668918006215.

[I 2021-09-28 20:04:15,496] Trial 93 finished with value: 0.7079562339592058 and parameters: {'booster': 'dart', 'learning\_rate': 0.011268334253279074, 'max\_depth': 6, 'subsample': 0.9267646452660486, 'colsample\_bytree': 0.8507211332081435}. Best is trial 48 with value: 0.7216668918006215.

[I 2021-09-28 20:04:15,803] Trial 94 finished with value: 0.696609482642172 and parameters: {'booster': 'dart', 'learning\_rate': 0.011823891214199193, 'max\_depth': 3, 'subsample': 0.9787513777651571, 'colsample\_bytree': 0.7981844754626292}. Best is trial 48 with value: 0.7216668918006215.

[I 2021-09-28 20:04:16,597] Trial 95 finished with value: 0.7180872619208428 and parameters: {'booster': 'dart', 'learning\_rate': 0.016977644190018425, 'max\_depth': 11, 'subsample': 0.8726989484903085, 'colsample\_bytree': 0.7750906228719172}. Best is trial 48 with value: 0.7216668918006215.

[I 2021-09-28 20:04:17,387] Trial 96 finished with value: 0.7208564095636903 and parameters: {'booster': 'dart', 'learning\_rate': 0.01048803022670973, 'max\_depth': 11, 'subsample': 0.8980363252192137, 'colsample\_bytree': 0.7259201370446458}. Best is trial 48 with value: 0.7216668918006215.

[I 2021-09-28 20:04:18,176] Trial 97 finished with value: 0.7206537890044576 and parameters: {'booster': 'dart', 'learning\_rate': 0.010417450718561778, 'max\_depth': 11, 'subsample': 0.8944598229534497, 'colsample\_bytree': 0.7221760100735997}. Best is trial 48 with value: 0.7216668918006215.

[I 2021-09-28 20:04:18,261] Trial 98 finished with value: 0.6946508172362555 and parameters: {'booster': 'gblinear', 'learning\_rate': 0.010398387418788381, 'max\_depth': 11, 'subsample': 0.9003173884353307, 'colsample\_bytree': 0.678602285997439}. Best is trial 48 with value: 0.7216668918006215.

[20:04:18] WARNING: C:/Users/Administrator/workspace/xgboost-win64\_release\_1.4.0/src/learner.cc:573:

Parameters: { "colsample\_bytree", "max\_depth", "subsample" } might not be used.

This may not be accurate due to some parameters are only used in language bindings but

passed down to XGBoost core. Or some parameters are not used but slip through this

verification. Please open an issue if you find above cases.

[I 2021-09-28 20:04:18,563] Trial 99 finished with value: 0.6983655274888558 and parameters: {'booster': 'dart', 'learning\_rate': 0.01462279862609719, 'max\_depth': 4, 'subsample': 0.8820371149015385, 'colsample\_bytree': 0.7237559712045788}. Best is trial 48 with value: 0.7216668918006215.

In [27]:

```
new_params = study.best_params

new_model = XGBClassifier(**new_params)
new_model.fit(X, y)
preds = new_model.predict(X_test)
```

```
print('Optimized SuperLearner accuracy: ', accuracy_score(y_test, preds))
print('Optimized SuperLearner f1-score: ', f1_score(y_test, preds, average='micro'))
```

C:\Users\burak\AppData\Roaming\Python\Python39\site-packages\xgboost\sklearn.py:1146: UserWarning: The use of label encoder in XGBClassifier is deprecated and will be removed in a future release. To remove this warning, do the following: 1) Pass option use\_label\_encoder=False when constructing XGBClassifier object; and 2) Encode your labels (y) as integers starting with 0, i.e. 0, 1, 2, ..., [num\_class - 1].

warnings.warn(label\_encoder\_deprecation\_msg, UserWarning)

[20:04:18] WARNING: C:/Users/Administrator/workspace/xgboost-win64\_release\_1.4.0/src/learner.cc:1095: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval\_metric if you'd like to restore the old behavior.

Optimized SuperLearner accuracy: 0.8114953397271376

Optimized SuperLearner f1-score: 0.8114953397271376

In [28]:

```
print("Number of finished trials: ", len(study.trials))
print("Best trial:")
trial = study.best_trial

print("  Value: {}".format(trial.value))
print("  Params: ")
for key, value in trial.params.items():
    print("    {}: {}".format(key, value))
```

Number of finished trials: 100

Best trial:

Value: 0.7216668918006215

Params:

booster: dart

learning\_rate: 0.012372918087737834

max\_depth: 11

subsample: 0.9655743598779992

colsample\_bytree: 0.7652990096612262

In [29]:

```
print("All of accuracies")
print(scores)

print("Mean of accuracies")
print(np.mean(scores))
```

All of accuracies

[0.7479015918958032, 0.7416413373860182, 0.7445361123172674, 0.7346938775510204, 0.7344044000578955]

Mean of accuracies

0.7406354638416011

In [30]:

```
def plot_confusion_matrix(cm, classes,
                          normalize=False,
                          title='Confusion matrix',
                          cmap=plt.cm.Blues):

    if normalize:
        cm = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]
        print("Normalized confusion matrix")
    else:
        print('Confusion matrix, without normalization')

    print(cm)

    plt.imshow(cm, interpolation='nearest', cmap=cmap)
    plt.title(title)
    plt.colorbar()
    tick_marks = np.arange(len(classes))
    plt.xticks(tick_marks, classes, rotation=45)
```

```
plt.yticks(tick_marks, classes)

fmt = '.2f' if normalize else 'd'
thresh = cm.max() / 2.
for i, j in itertools.product(range(cm.shape[0]), range(cm.shape[1])):
    plt.text(j, i, format(cm[i, j], fmt),
             horizontalalignment="center",
             color="white" if cm[i, j] > thresh else "black")

plt.tight_layout()
plt.ylabel('True label')
plt.xlabel('Predicted label')
```

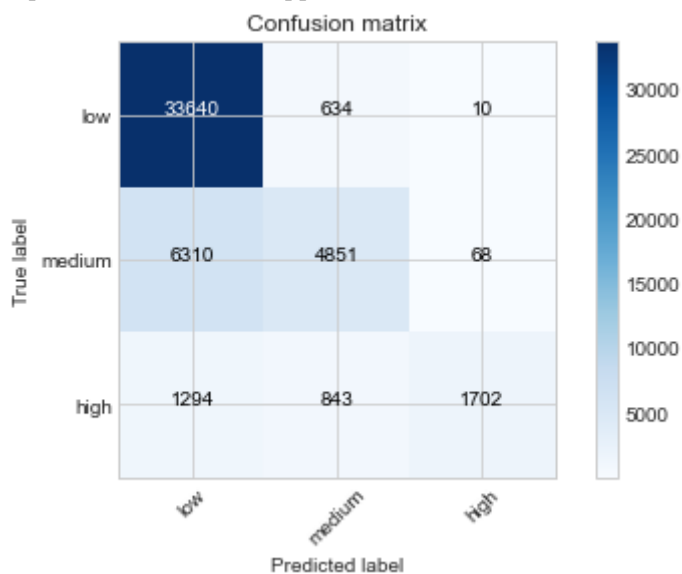
In [31]:

```
y_pred = new_model.predict(X)
cm = confusion_matrix(train_df['target'], y_pred)
np.set_printoptions(precision=2)

class_names = ['low', 'medium', 'high']
# Plot non-normalized confusion matrix
plt.figure()
plot_confusion_matrix(cm, classes=class_names,
                      title='Confusion matrix')
```

Confusion matrix, without normalization

```
[[33640  634   10]
 [ 6310 4851   68]
 [ 1294  843 1702]]
```



## FEATURE IMPORTANCE BY LOFO

In [32]:

```
# define the validation scheme
cv = KFold(n_splits=4, shuffle=True, random_state=0)
scorer = make_scorer(mean_absolute_error, greater_is_better=False)
# define the binary target and the features
target = "target"
features = [col for col in train_df.columns if col != target]
dataset = Dataset(df=train_df, target=target, features=features)
# define the validation scheme and scorer. The default model is LightGBM
lofo_imp = LOFOImportance(dataset, scoring=scorer, model=new_model, cv=cv)

# get the mean and standard deviation of the importances in pandas format
importance_df = lofo_imp.get_importance()
```

```
# plot the means and standard deviations of the importances
plot_importance(importance_df)
```

```
[20:05:15] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.4.0/src/learner.cc:1095: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[20:05:55] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.4.0/src/learner.cc:1095: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[20:06:35] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.4.0/src/learner.cc:1095: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[20:07:15] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.4.0/src/learner.cc:1095: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[20:07:57] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.4.0/src/learner.cc:1095: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[20:08:37] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.4.0/src/learner.cc:1095: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[20:09:17] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.4.0/src/learner.cc:1095: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[20:10:00] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.4.0/src/learner.cc:1095: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[20:10:44] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.4.0/src/learner.cc:1095: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[20:11:25] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.4.0/src/learner.cc:1095: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[20:12:06] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.4.0/src/learner.cc:1095: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[20:12:45] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.4.0/src/learner.cc:1095: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[20:13:26] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.4.0/src/learner.cc:1095: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[20:14:07] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.4.0/src/learner.cc:1095: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[20:14:55] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.4.0/src/learner.cc:1095: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[20:15:37] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.4.0/src/learner.cc:1095: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[20:16:18] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.4.0/src
```



[illegible]

[illegible]

[illegible]

```
[20:58:51] WARNING: C:/Users/Administrator/workspace/xgboost-win64 release 1.4.0/sr
```

[illegible]

