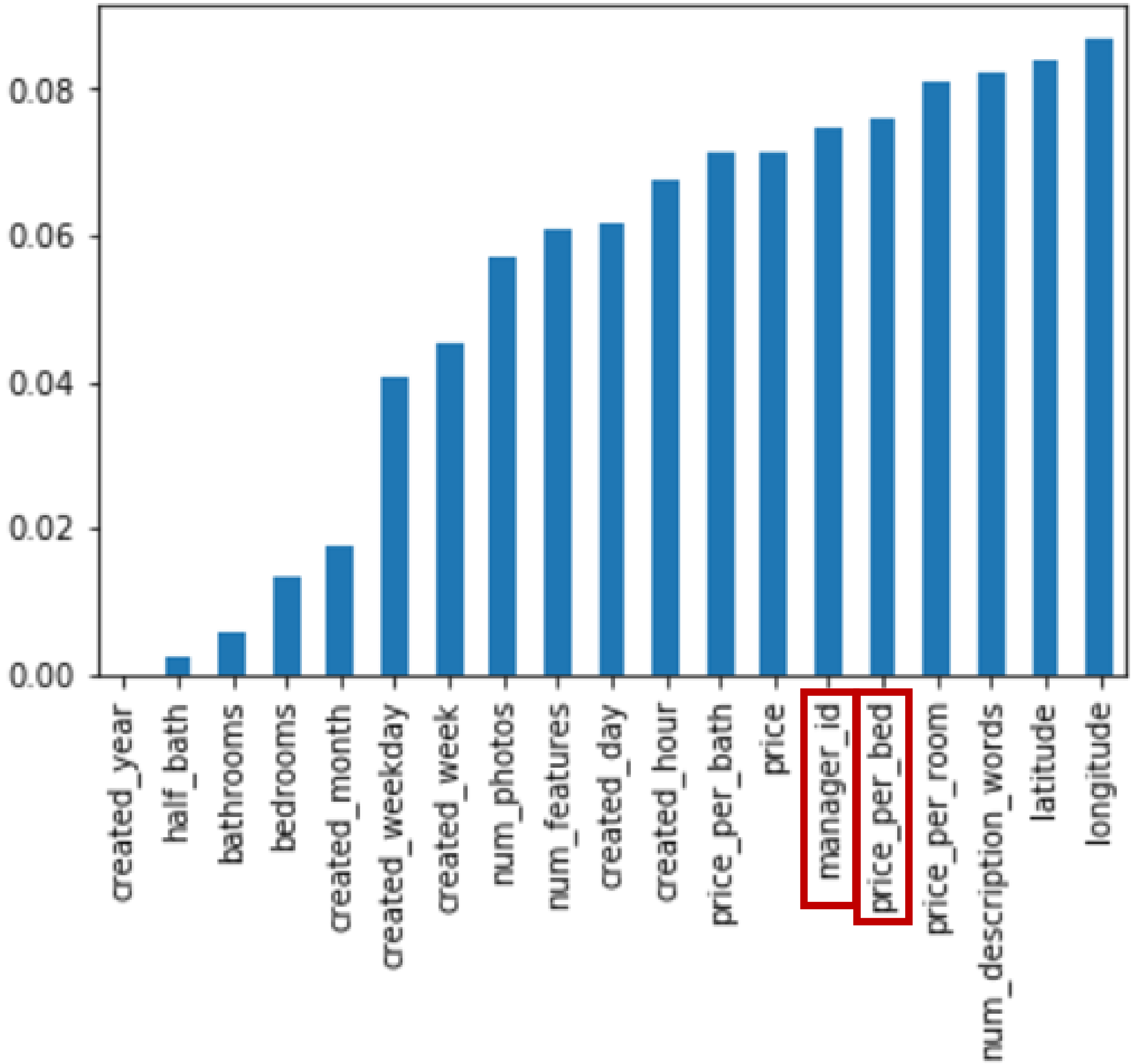
Kaggle #2
Feature Engineering

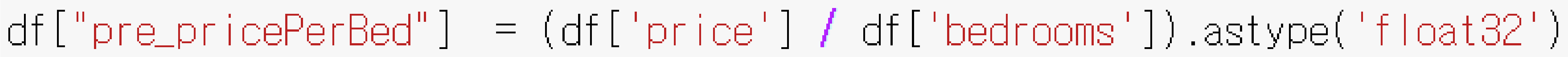
서 예지
장 예훈
조 용걸

1. Feature Selection



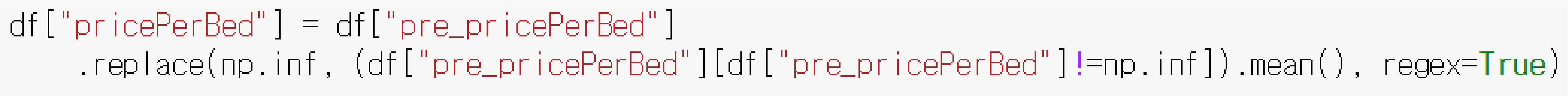
1. Feature Selection

- Price per Bed



```
df["pre_pricePerBed"]  = (df['price'] / df['bedrooms']).astype('float32')
```

(가격 / 침대 개수)

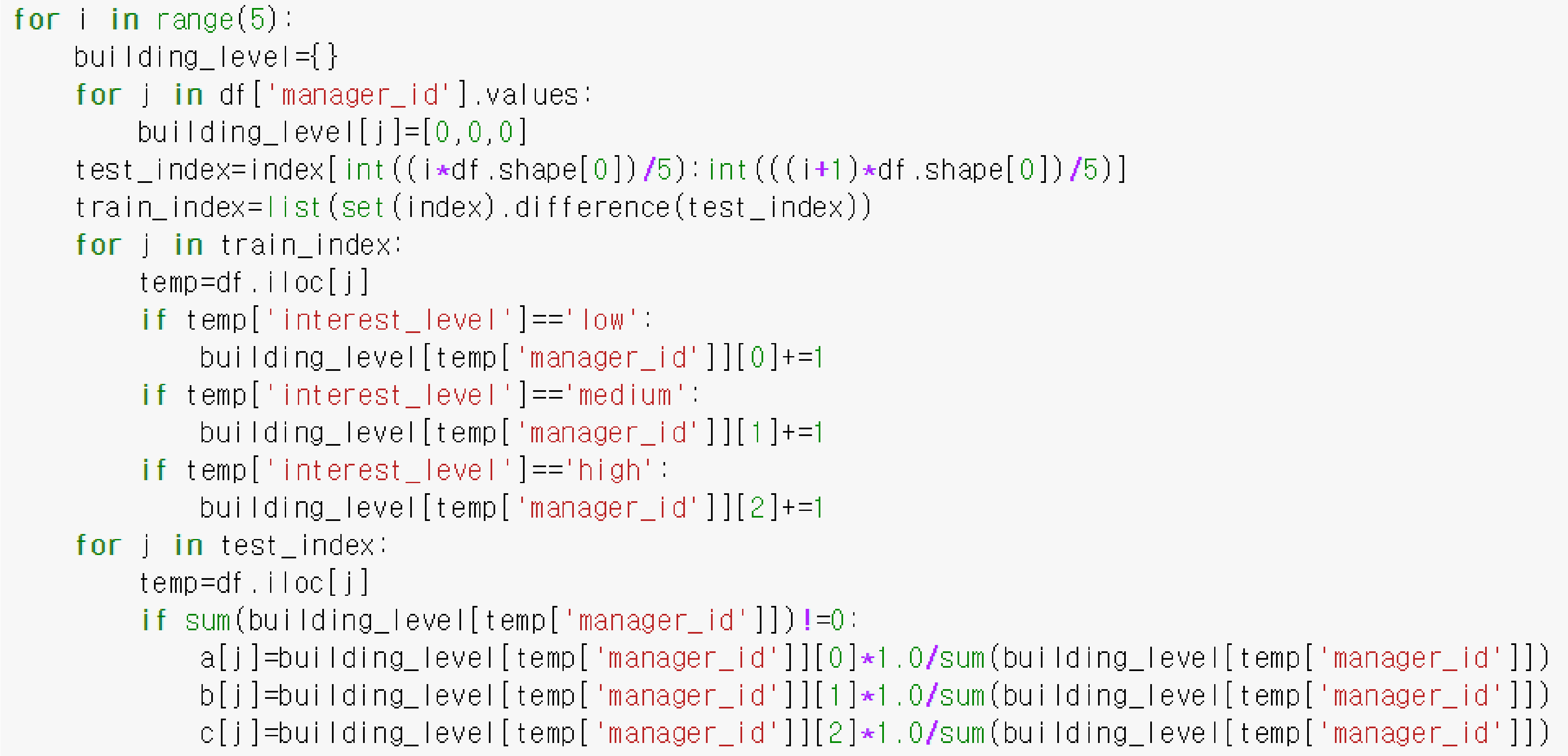


```
df["pricePerBed"] = df["pre_pricePerBed"]
    .replace(np.inf, (df["pre_pricePerBed"][df["pre_pricePerBed"]!=np.inf]).mean(), regex=True)
```

Inf값을 “PricePerBed” 의 평균 값으로 대체

1. Feature Selection

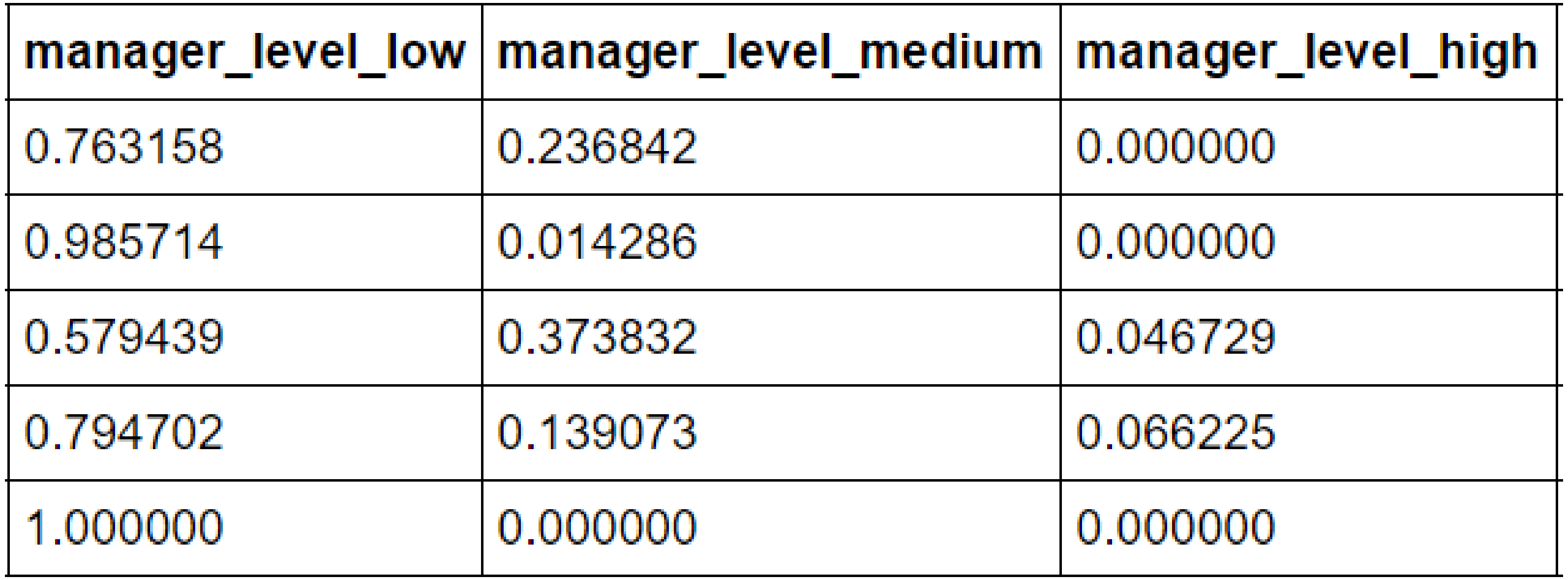
- Manager Id



```
for i in range(5):
    building_level={}
    for j in df['manager_id'].values:
        building_level[j]=[0,0,0]
    test_index=index[int((i*df.shape[0])/5):int(((i+1)*df.shape[0])/5)]
    train_index=list(set(index).difference(test_index))
    for j in train_index:
        temp=df.iloc[j]
        if temp['interest_level']=='low':
            building_level[temp['manager_id']][0]+=1
        if temp['interest_level']=='medium':
            building_level[temp['manager_id']][1]+=1
        if temp['interest_level']=='high':
            building_level[temp['manager_id']][2]+=1
    for j in test_index:
        temp=df.iloc[j]
        if sum(building_level[temp['manager_id']])!=0:
            a[j]=building_level[temp['manager_id']][0]*1.0/sum(building_level[temp['manager_id']])
            b[j]=building_level[temp['manager_id']][1]*1.0/sum(building_level[temp['manager_id']])
            c[j]=building_level[temp['manager_id']][2]*1.0/sum(building_level[temp['manager_id']])
```

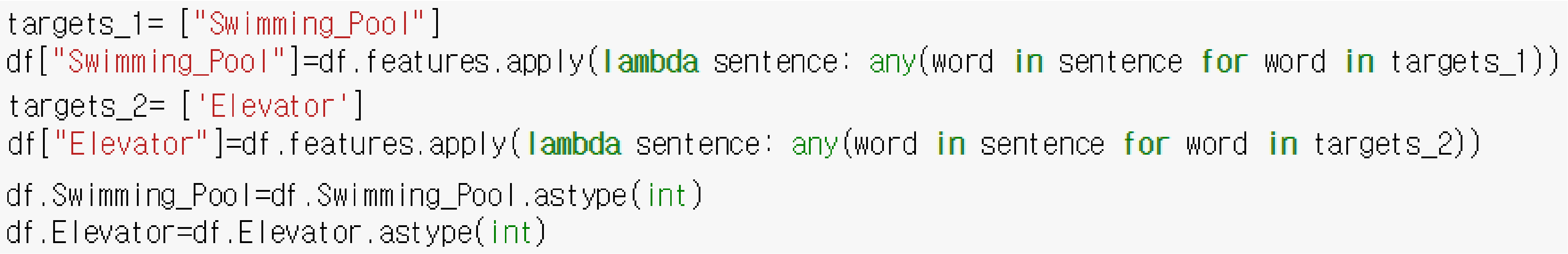
Manager_id를 [“High” , “Midum” , “Low”]로 Level 구분

1. Feature Selection

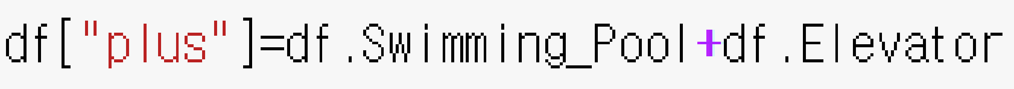


manager_level_low	manager_level_medium	manager_level_high
0.763158	0.236842	0.000000
0.985714	0.014286	0.000000
0.579439	0.373832	0.046729
0.794702	0.139073	0.066225
1.000000	0.000000	0.000000

1. Feature Selection



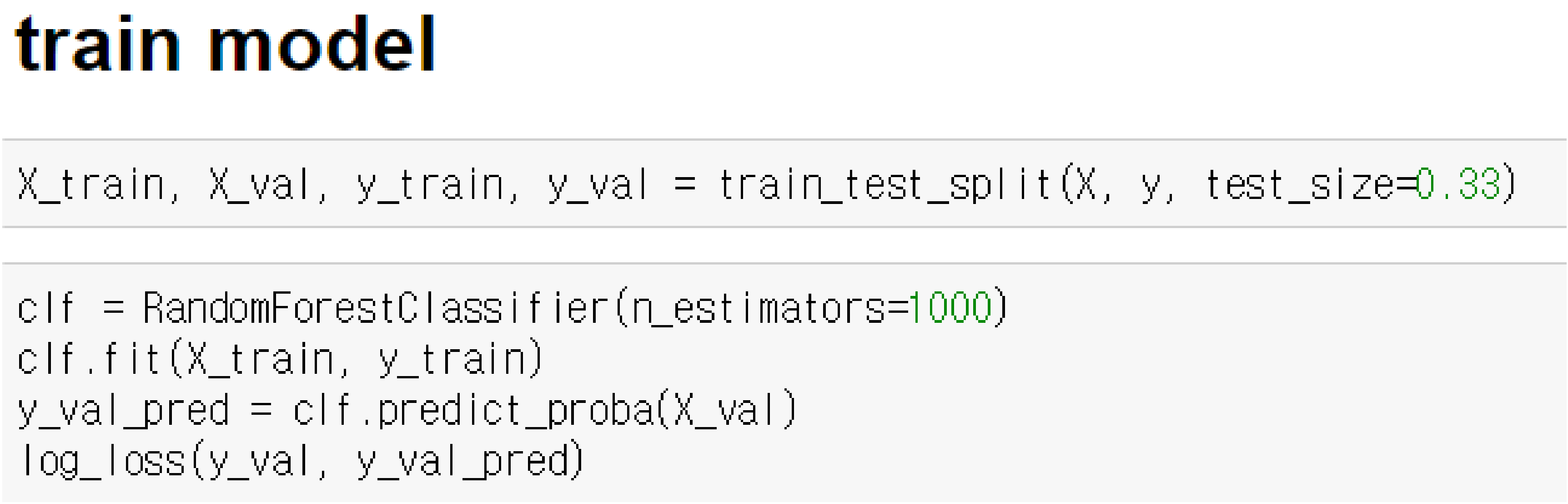
```
targets_1= ["Swimming_Pool"]
df["Swimming_Pool"]=df.features.apply(lambda sentence: any(word in sentence for word in targets_1))
targets_2= ['Elevator']
df["Elevator"]=df.features.apply(lambda sentence: any(word in sentence for word in targets_2))
df.Swimming_Pool=df.Swimming_Pool.astype(int)
df.Elevator=df.Elevator.astype(int)
```



```
df["plus"]=df.Swimming_Pool+df.Elevator
```

Features Data 중 "Swimming_Pool" 과 "Elevator" 를 합친 "Plus" 생성

2. LogLoss



train model

```
X_train, X_val, y_train, y_val = train_test_split(X, y, test_size=0.33)
```

```
clf = RandomForestClassifier(n_estimators=1000)
clf.fit(X_train, y_train)
y_val_pred = clf.predict_proba(X_val)
log_loss(y_val, y_val_pred)
```

LogLoss = 0.5864