A4_1914068_ID3_code

March 28, 2022

[1]:
```
import pandas as pd
from sklearn import tree
import matplotlib.pyplot as plt
```

[2]:
```
df = pd.read_csv("A4_1914068_train.csv")
```

[3]:
```
features = ['LotArea', '1stFlrSF', 'FullBath', 'TotRmsAbvGrd', 'SalePrice']
df_final = df[features].copy()
df_final.dropna(inplace=True)
```

[4]:
```
df_final.head()
df_final = df_final[:100]
```

[5]:
```
df_final["SalePriceNorm"] = df.apply(lambda x: (x['SalePrice'] -␣
↪df_final['SalePrice'].min())/(df_final['SalePrice'].max() -␣
↪df_final['SalePrice'].min()), axis=1)
```

[6]:
```
def categorize(var):
    if var < 1/3:
        return "Low"
    elif var >= 1/3 and var < 2/3:
        return "Medium"
    else:
        return "High"
```

[7]:
```
df_final["SalePriceCategory"] = df_final.apply(lambda x:␣
↪categorize(x["SalePriceNorm"]), axis = 1)
```

[8]:
```
df_final.head()
```

[8]:

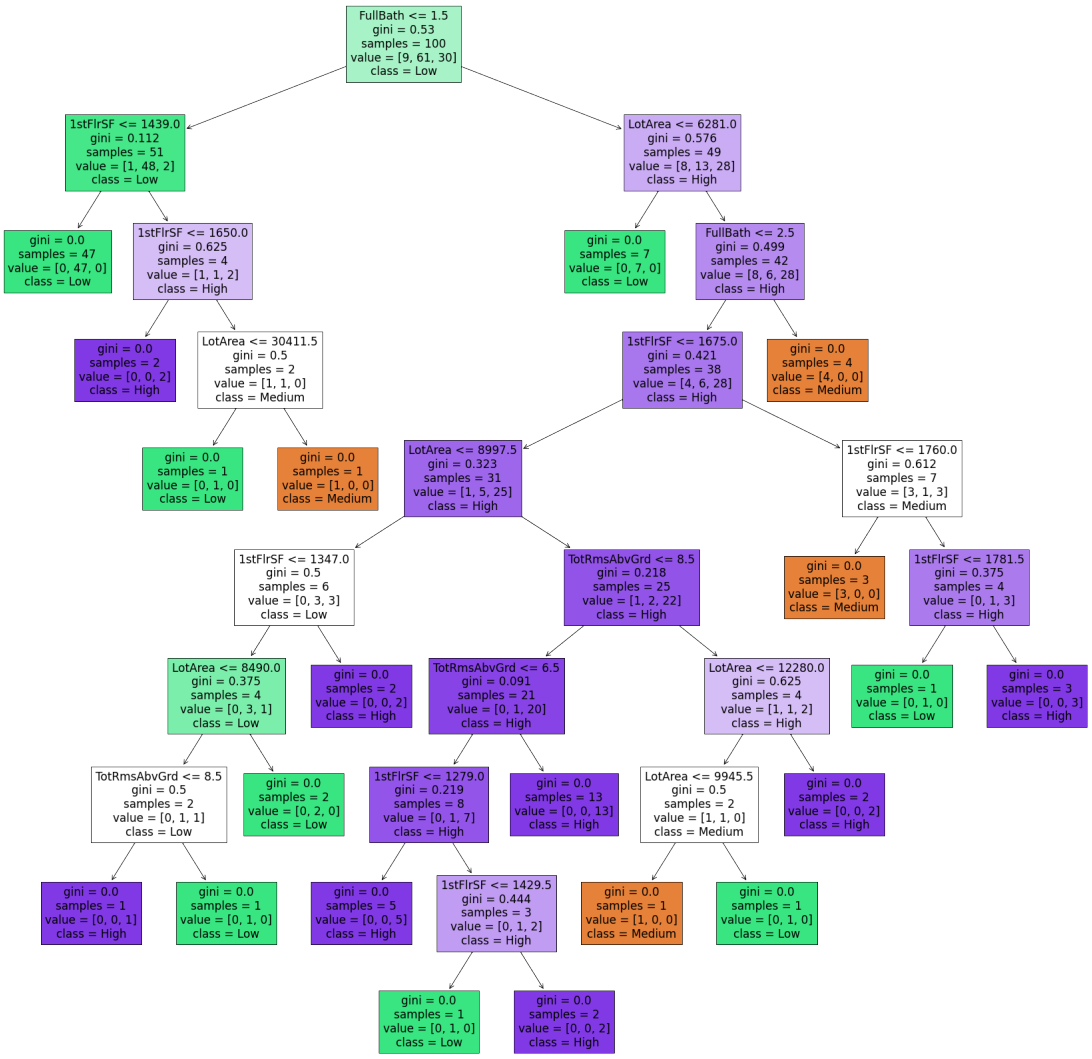
	LotArea	1stFlrSF	FullBath	TotRmsAbvGrd	SalePrice	SalePriceNorm	\
0	8450	856	2	8	208500	0.422539	
1	9600	1262	2	6	181500	0.354832	
2	11250	920	2	6	223500	0.460153	
3	9550	961	1	7	140000	0.250765	
4	14260	1145	2	9	250000	0.526606	

```
  SalePriceCategory
0            Medium
1            Medium
2            Medium
3               Low
4            Medium
```

[9]:
```
X, y = df_final[['LotArea', '1stFlrSF', 'FullBath', 'TotRmsAbvGrd']],␣
↪df_final['SalePriceCategory']
```

[10]:
```
clf = tree.DecisionTreeClassifier()
clf = clf.fit(X, y)
```

[11]:
```
plt.figure(figsize=(30, 30)) # Resize figure
tree.plot_tree(clf, filled=True,
               feature_names= ['LotArea', '1stFlrSF', 'FullBath',␣
↪'TotRmsAbvGrd', 'SalePrice'],
              class_names = df_final['SalePriceCategory'].unique())
plt.savefig("tree.png")
plt.show()
```


FullBath <= 1.5
gini = 0.53
samples = 100
value = [9, 61, 30]
class = Low
1stFlrSF <= 1439.0
gini = 0.112
samples = 51
value = [1, 48, 2]
class = Low
LotArea <= 6281.0
gini = 0.576
samples = 49
value = [8, 13, 28]
class = High
gini = 0.0
samples = 47
value = [0, 47, 0]
class = Low
1stFlrSF <= 1650.0
gini = 0.625
samples = 4
value = [1, 1, 2]
class = High
gini = 0.0
samples = 7
value = [0, 7, 0]
class = Low
FullBath <= 2.5
gini = 0.499
samples = 42
value = [8, 6, 28]
class = High
gini = 0.0
samples = 2
value = [0, 0, 2]
class = High
LotArea <= 30411.5
gini = 0.5
samples = 2
value = [1, 1, 0]
class = Medium
1stFlrSF <= 1675.0
gini = 0.421
samples = 38
value = [4, 6, 28]
class = High
gini = 0.0
samples = 4
value = [4, 0, 0]
class = Medium
gini = 0.0
samples = 1
value = [0, 1, 0]
class = Low
gini = 0.0
samples = 1
value = [1, 0, 0]
class = Medium
LotArea <= 8997.5
gini = 0.323
samples = 31
value = [1, 5, 25]
class = High
1stFlrSF <= 1760.0
gini = 0.612
samples = 7
value = [3, 1, 3]
class = Medium
1stFlrSF <= 1347.0
gini = 0.5
samples = 6
value = [0, 3, 3]
class = Low
TotRmsAbvGrd <= 8.5
gini = 0.218
samples = 25
value = [1, 2, 22]
class = High
gini = 0.0
samples = 3
value = [3, 0, 0]
class = Medium
1stFlrSF <= 1781.5
gini = 0.375
samples = 4
value = [0, 1, 3]
class = High
LotArea <= 8490.0
gini = 0.375
samples = 4
value = [0, 3, 1]
class = Low
gini = 0.0
samples = 2
value = [0, 0, 2]
class = High
TotRmsAbvGrd <= 6.5
gini = 0.091
samples = 21
value = [0, 1, 20]
class = High
LotArea <= 12280.0
gini = 0.625
samples = 4
value = [1, 1, 2]
class = High
gini = 0.0
samples = 1
value = [0, 1, 0]
class = Low
gini = 0.0
samples = 3
value = [0, 0, 3]
class = High
TotRmsAbvGrd <= 8.5
gini = 0.5
samples = 2
value = [0, 1, 1]
class = Low
gini = 0.0
samples = 2
value = [0, 2, 0]
class = Low
1stFlrSF <= 1279.0
gini = 0.219
samples = 8
value = [0, 1, 7]
class = High
gini = 0.0
samples = 13
value = [0, 0, 13]
class = High
LotArea <= 9945.5
gini = 0.5
samples = 2
value = [1, 1, 0]
class = Medium
gini = 0.0
samples = 2
value = [0, 0, 2]
class = High
gini = 0.0
samples = 1
value = [0, 0, 1]
class = High
gini = 0.0
samples = 1
value = [0, 1, 0]
class = Low
gini = 0.0
samples = 5
value = [0, 0, 5]
class = High
1stFlrSF <= 1429.5
gini = 0.444
samples = 3
value = [0, 1, 2]
class = High
gini = 0.0
samples = 1
value = [1, 0, 0]
class = Medium
gini = 0.0
samples = 1
value = [0, 1, 0]
class = Low
gini = 0.0
samples = 1
value = [0, 1, 0]
class = Low
gini = 0.0
samples = 2
value = [0, 0, 2]
class = High

[ ]: