project4_Jacob_Padgett.md

Project 4 - Jacob Padgett

Elevator pitch

The state of Colorado has a large portion of their residential dwelling data that is missing the year built and they would like you to a predictive model that can classify if a house is built pre 1980. They would also like you to build a model that predicts (regression actual age of each home.

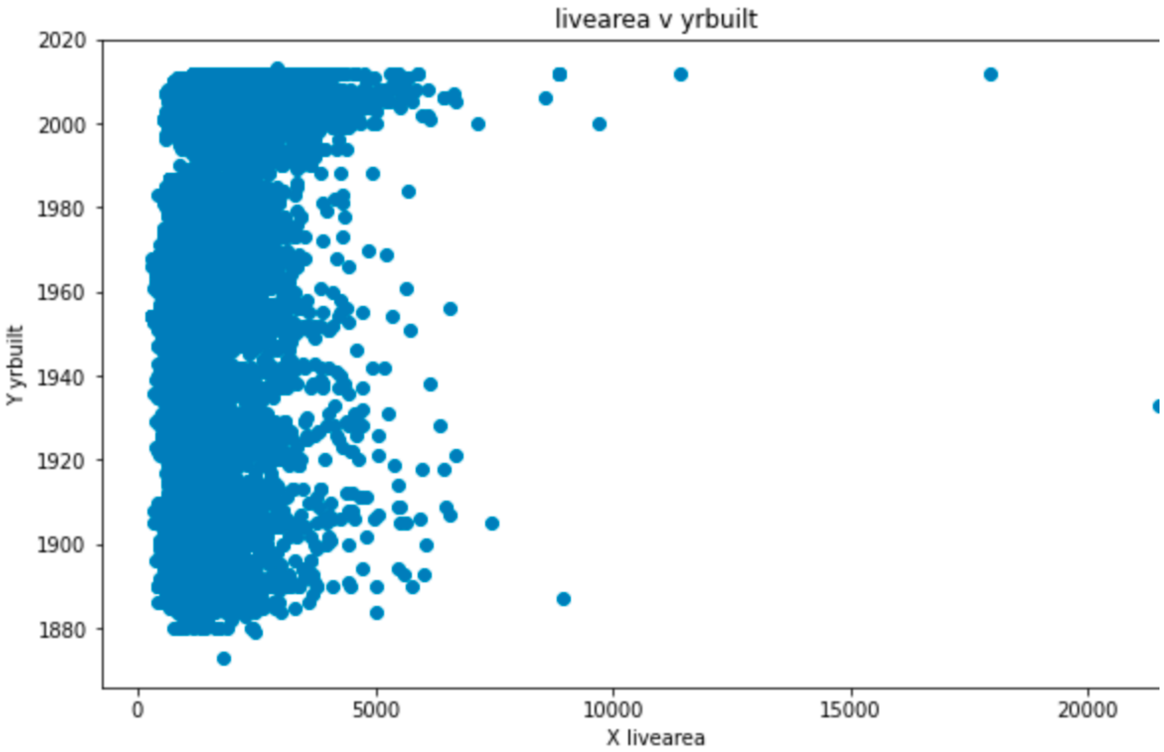
TECHNICAL DETAILS

I was told not to repeat myself, so I'll skip this section as everything will be included below.

GRAND QUESTION 1

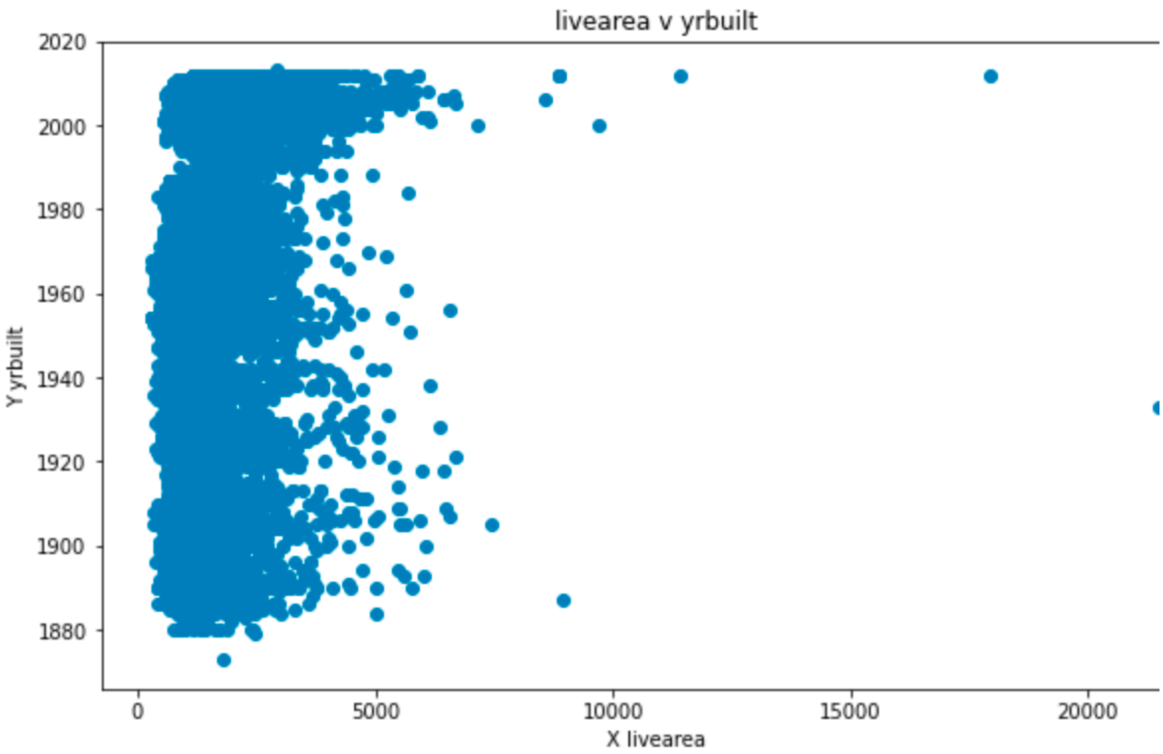
Create 2-3 charts that evaluate potential relationships between the home variables and before1980.

- Chart 1 is livearea v yrbuilt



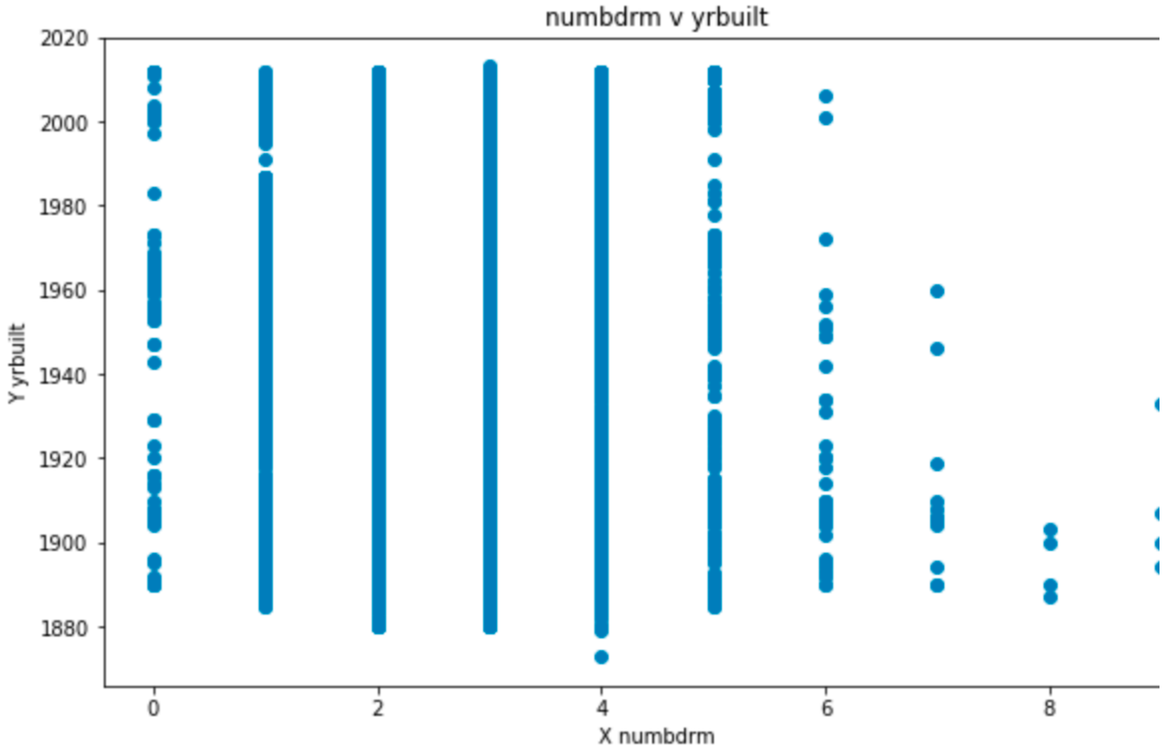
    ○

- Chart 2 is numbdrm v yrbuilt



- ○

- Chart 3 is basement v yrbuilt



- ○

GRAND QUESTION 2

Can you build a classification model (before or after 1980) that has at least 90% accuracy for the state of Colorado to use (e your model choice and which models you tried)?

```
# Removes the target and keeps all features
X = df2.drop('yrbuilt', axis=1)

# Selects the target column
y = df2['yrbuilt']

# Splitting X and y variables into train and test sets using stratified sampling
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
                                                    random_state=24)

# Creating random forest object
rf = RandomForestClassifier(random_state=24)

# Fit with the training data
rf.fit(X_train, y_train)

# Using the features in the test set to make predictions
y_pred = rf.predict(X_test)

# Comparing predictions to actual values
print(f'accuracy_score = {accuracy_score(y_test, y_pred)}')
```

The result showed it had an `accuracy_score = 0.2830957230142566` . Not quite 90%.

GRAND QUESTION 3

Will you justify your classification model by detailing the most important features in your model (a chart and a description ar must)?

The following code provides me my answers for this section showing livearea has the largest influence:

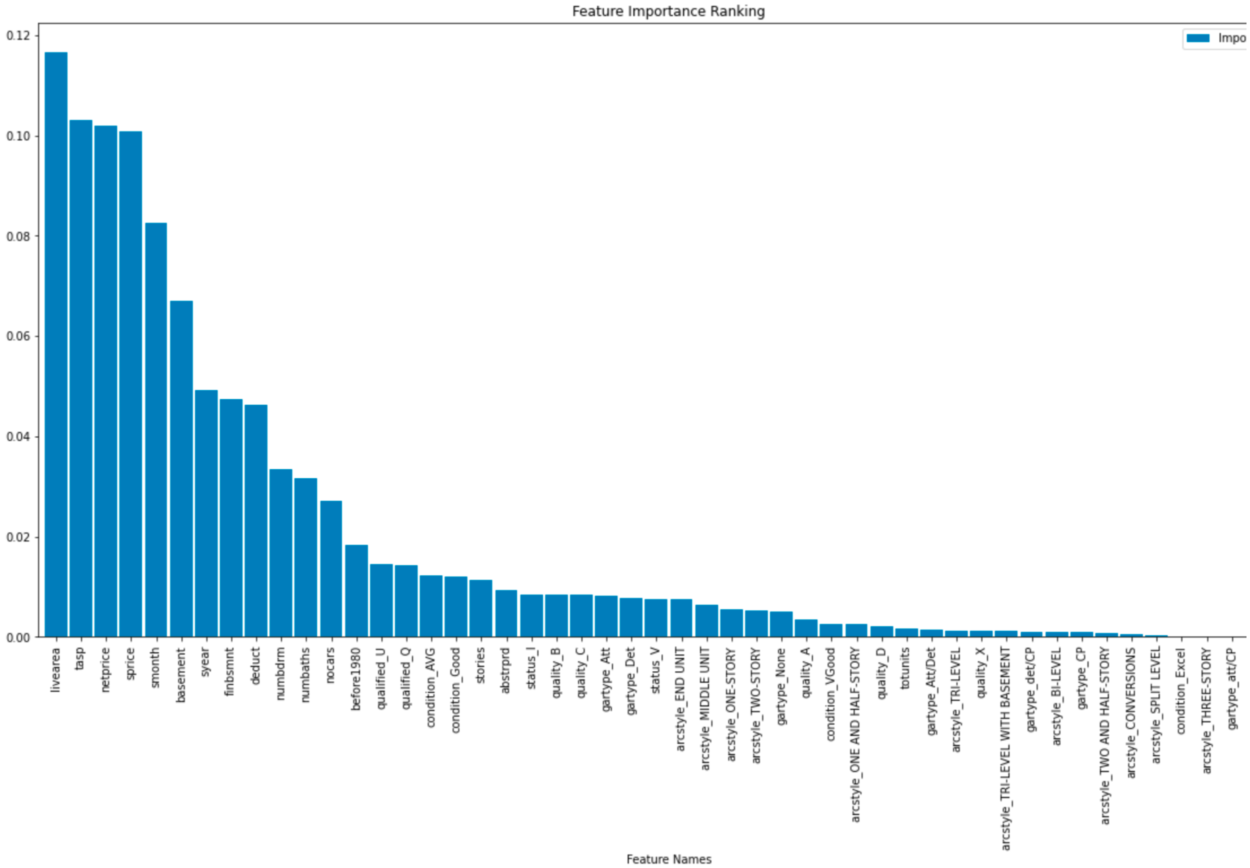
```
feat_imports = (pd.DataFrame(
    {"Feature Names": X_train.columns,
    "Importances": rf.feature_importances_})
    .sort_values("Importances", ascending=False))

print(feat_imports.to_markdown(index=False))

feat_imports.plot.bar(x='Feature Names', y='Importances', rot=90, width=.9,figsize=(20,10), title="Feature
Importance Ranking")
```

Here is the table followed by the chart:

Feature Names	Importances
livearea	0.116623
tasp	0.10299
netprice	0.101838
sprice	0.100706
smonth	0.0824698
basement	0.0670488
syear	0.0491019
finbsmnt	0.047474

Feature Names	Importances
deduct	0.0463834
numbdrm	0.0334988
numbaths	0.0315304
nocars	0.0272065
before1980	0.0183445
qualified_U	0.0146188
qualified_Q	0.0141859
condition_AVG	0.0123554
condition_Good	0.0120149
stories	0.0113472
abstrprd	0.00923042
status_I	0.0084916
quality_B	0.00838169
quality_C	0.00833747
gartype_Att	0.00822686
gartype_Det	0.00779408
status_V	0.00757432
arcstyle_END UNIT	0.00745558
arcstyle_MIDDLE UNIT	0.00629261
arcstyle_ONE-STORY	0.00551416
arcstyle_TWO-STORY	0.00525629
gartype_None	0.00509823
quality_A	0.0035479
condition_VGood	0.00263312
arcstyle_ONE AND HALF-STORY	0.00249928
quality_D	0.00208182
totunits	0.00173194
gartype_Att/Det	0.00142104
arcstyle_TRI-LEVEL	0.0013099
quality_X	0.00130661
arcstyle_TRI-LEVEL WITH BASEMENT	0.00111844
gartype_det/CP	0.0010197
arcstyle_BI-LEVEL	0.000926003
gartype_CP	0.000917346

Feature Names	Importances
arcstyle_TWO AND HALF-STORY	0.000680208
arcstyle_CONVERSIONS	0.000600348
arcstyle_SPLIT LEVEL	0.000263447
condition_Excel	0.000184948
arcstyle_THREE-STORY	0.000183382
gartype_att/CP	0.000150588
condition_Fair	3.24045e-05
condition_None	0



GRAND QUESTION 4

Can you describe the quality of your classification model using 2-3 evaluation metrics? You need to provide an interpretation each evaluation metric when you provide the value.

```
answer = """
The question seems to be asking for a yes/no answer. So, no. I can not.

Please provide feedback that is useful and offers instruction/examples of what is supposed to happen with this
assignment as I'm sure I got everything wrong. I look forward to learning more about ML.
"""
print(answer)
```

APPENDIX A (PYTHON SCRIPT)

```
# %%
# Loading in packages
from re import T
import pandas as pd
import numpy as np
import altair as alt
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
import seaborn as sns
import matplotlib.pyplot as plt
# %%
# Grand Questions:

# Create 2-3 charts that evaluate potential relationships between the home variables and before1980.

# Can you build a classification model (before or after 1980) that has at least 90% accuracy for the state of
Colorado to use (explain your model choice and which models you tried)?

# Will you justify your classification model by detailing the most important features in your model (a chart and
description are a must)?

# Can you describe the quality of your classification model using 2-3 evaluation metrics? You need to provide an
interpretation of each evaluation metric when you provide the value.
# %%
# read in data
url1 = 'https://github.com/byuidatascience/data4dwellings/raw/master/data-
raw/dwellings_denver/dwellings_denver.csv'
url2 = 'https://github.com/byuidatascience/data4dwellings/raw/master/data-raw/dwellings_ml/dwellings_ml.csv'
url3 = 'https://github.com/byuidatascience/data4dwellings/raw/master/data-
raw/dwellings_neighborhoods_ml/dwellings_neighborhoods_ml.csv'

df1 = pd.read_csv(url1) # load first set
df1.drop('xtraffic', axis=1, inplace=True) # drop what seems to be a useless col

df2 = pd.read_csv(url2) # load second set

df3 = pd.read_csv(url3) # load third set
# %%
# Create 2-3 charts that evaluate potential relationships between the home variables and before1980.

# Set the figure size in inches
plt.figure(figsize=(10,6))

# plt.scatter(x, y, label = "label_name" )
plt.scatter(df2['livearea'], df2['yrbuilt'])#, label = "livearea v yrbuilt" )

# Set x and y axes labels
plt.xlabel('X livearea')
plt.ylabel('Y yrbuilt')

plt.title('livearea v yrbuilt')
plt.legend()
plt.show()
# %%
# Set the figure size in inches
plt.figure(figsize=(10,6))

# plt.scatter(x, y, label = "label_name" )
plt.scatter(df2['numbdrm'], df2['yrbuilt'])#, label = "numbdrm v yrbuilt" )

# Set x and y axes labels
```

```
plt.xlabel('X numbdrm')
plt.ylabel('Y yrbuilt')

plt.title('numbdrm v yrbuilt')
plt.legend()
plt.show()

# %%
# Set the figure size in inches
plt.figure(figsize=(10,6))

# plt.scatter(x, y, label = "label_name" )
plt.scatter(df2['basement'], df2['yrbuilt'])#, label = "basement v yrbuilt" )

# Set x and y axes labels
plt.xlabel('X basement')
plt.ylabel('Y yrbuilt')

plt.title('basement v yrbuilt')
plt.legend()
plt.show()
# %%
# %%
# %%
# %%
# %%
# %%
# Can you build a classification model (before or after 1980) that has at least 90% accuracy for the state of
Colorado to use (explain your model choice and which models you tried)?

# Removes the target and keeps all features
X = df2.drop('yrbuilt', axis=1)

# Selects the target column
y = df2['yrbuilt']

# Splitting X and y variables into train and test sets using stratified sampling
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
                                                     random_state=24)#, stratify=y)

# Creating random forest object
rf = RandomForestClassifier(random_state=24)

# Fit with the training data
rf.fit(X_train, y_train)

# Using the features in the test set to make predictions
y_pred = rf.predict(X_test)

# Comparing predictions to actual values
print(f'accuracy_score = {accuracy_score(y_test, y_pred)}')
# %%
# %%
# %%
# %%
# %%
# %%
# Will you justify your classification model by detailing the most important features in your model (a chart and
description are a must)?

feat_imports = (pd.DataFrame(
    {"Feature Names": X_train.columns,
    "Importances": rf.feature_importances_})
    .sort_values("Importances", ascending=False))

print(feat_imports.to_markdown(index=False))
# %%
feat_imports.plot.bar(x='Feature Names', y='Importances', rot=90, width=.9,figsize=(20,10), title="Feature
Importance Ranking")
# %%
# %%
# %%
# %%
# %%
# %%
# Can you describe the quality of your classification model using 2–3 evaluation metrics? You need to provide an
interpretation of each evaluation metric when you provide the value.

answer = """
The question seems to be asking for a yes/no answer. So, no. I can not.

Please provide feedback that is useful and offers instruction/examples of what is supposed to happen with this
assignment as I'm sure I got everything wrong. I look forward to learning more about ML.
"""
print(answer)
# %%
```