Assignment 04

Forecast Sales

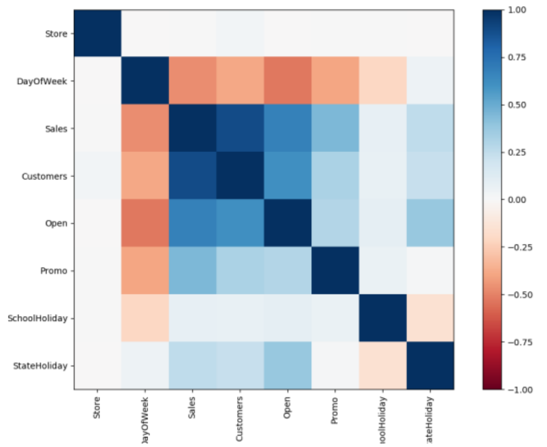
15146314 Yang, Seunghyuck

Select explanatory variables

Select learning methods

Summarize

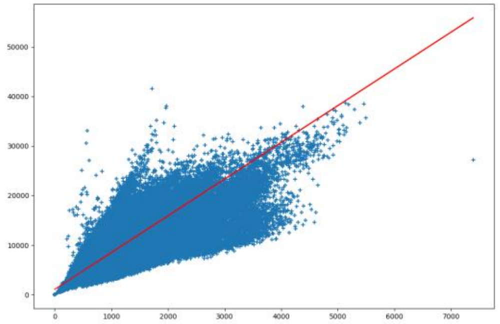
Correlation Graph

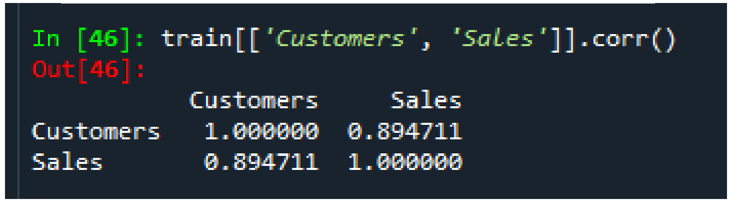


Tier 1: Customers, Open, Promo

Tier 2: SchoolHoliday, StateHoliday, DayOfWeek

Customer





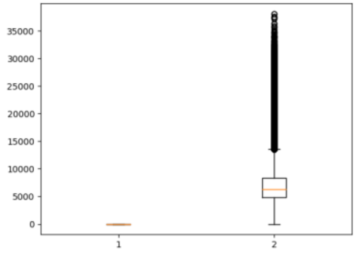
```
In [46]: train[['Customers', 'Sales']].corr()
Out[46]:
           Customers     Sales
Customers   1.000000  0.894711
Sales       0.894711  1.000000
```

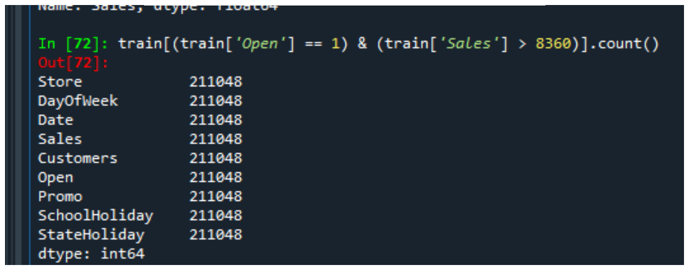
High correlation

Observable data

Select

Open





```
Name: Sales, dtype: float64

In [72]: train[(train['Open'] == 1) & (train['Sales'] > 8360)].count()
Out[72]:
Store            211048
DayOfWeek        211048
Date             211048
Sales            211048
Customers        211048
Open             211048
Promo            211048
SchoolHoliday    211048
StateHoliday     211048
dtype: int64
```

Binary data

Clear relationship

Reasonable correlation

Outliner is too lot

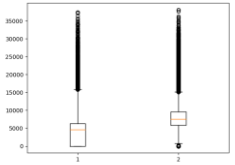
Select without transform

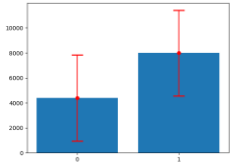
Select explanatory variables

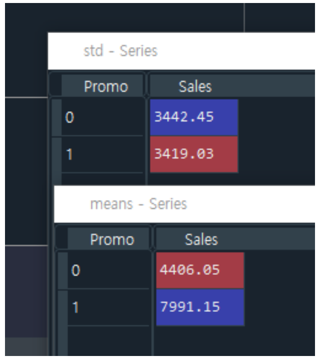
Select learning methods

Summarize

Promo







std - Series

Promo	Sales
0	3442.45
1	3419.03

means - Series

Promo	Sales
0	4406.05
1	7991.15

Binary data

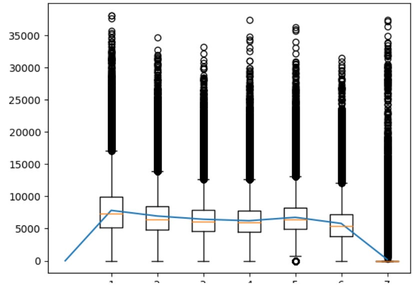
Prominent correlation

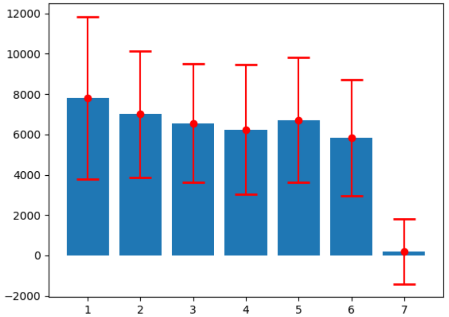
Observable gap

Standard deviation is too high

Discard

DayOfWeek





Categorical data

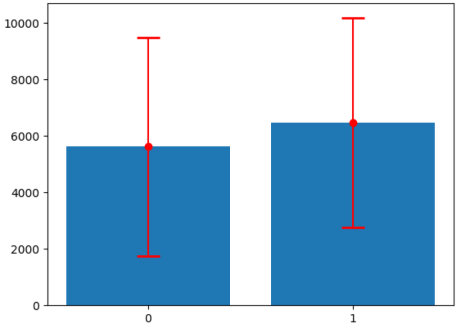
Low correlation

Barely see difference in overall

Clear gap between 7 and others

Select with transform

SchoolHoliday



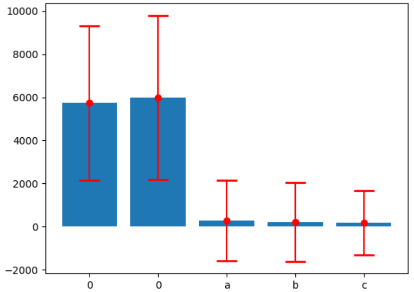
Binary data

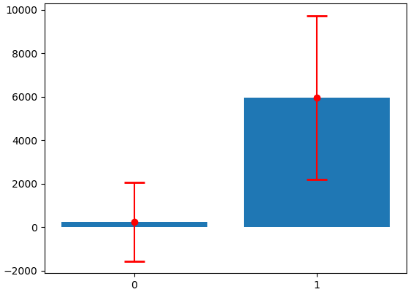
Low correlation

Barely see difference in overall

Discard

StateHoliday





Categorical data

Low correlation

Barely see difference in overall

Low correlation after transform

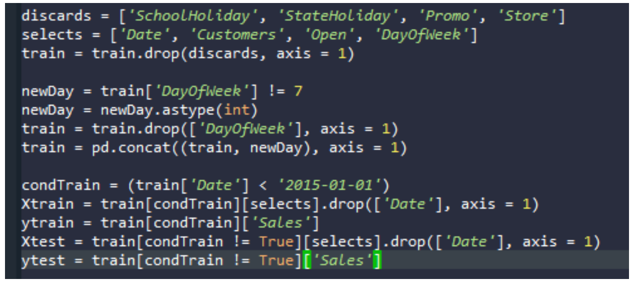
Discard

Select explanatory variables

Select learning methods

Summarize

Preparing data



```
discards = ['SchoolHoliday', 'StateHoliday', 'Promo', 'Store']
selects = ['Date', 'Customers', 'Open', 'DayOfWeek']
train = train.drop(discards, axis = 1)

newDay = train['DayOfWeek'] != 7
newDay = newDay.astype(int)
train = train.drop(['DayOfWeek'], axis = 1)
train = pd.concat((train, newDay), axis = 1)

condTrain = (train['Date'] < '2015-01-01')
Xtrain = train[condTrain][selects].drop(['Date'], axis = 1)
ytrain = train[condTrain]['Sales']
Xtest = train[condTrain != True][selects].drop(['Date'], axis = 1)
ytest = train[condTrain != True]['Sales']
```

Transforming data

Split data into 4 pieces

2013 – 2014 / 2015

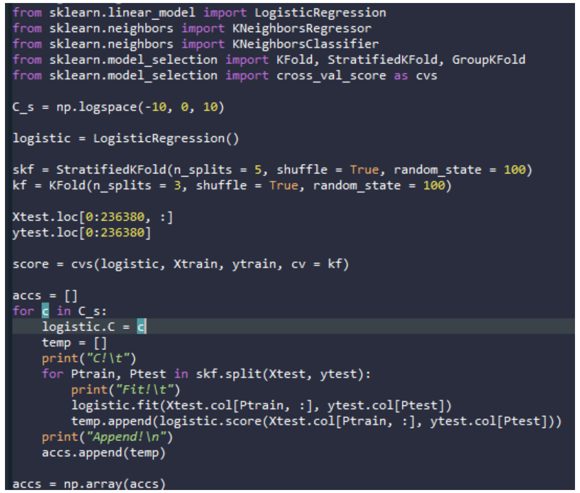
Xtrain, ytrain, Xtest, ytest

Evaluating variables through Cross Validation

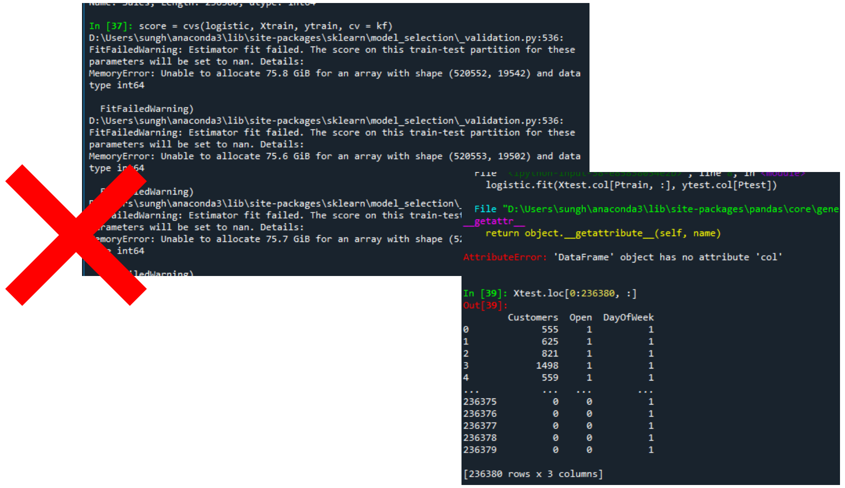
Logistic Regression

KFold

StratifiedKFold



```
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsRegressor
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import KFold, StratifiedKFold, GroupKFold
from sklearn.model_selection import cross_val_score as cvs

C_s = np.logspace(-10, 0, 10)

logistic = LogisticRegression()

skf = StratifiedKFold(n_splits = 5, shuffle = True, random_state = 100)
kf = KFold(n_splits = 3, shuffle = True, random_state = 100)

Xtest.loc[0:236380, :]
ytest.loc[0:236380]

score = cvs(logistic, Xtrain, ytrain, cv = kf)

accs = []
for c in C_s:
    logistic.C = c
    temp = []
    print("C!\t")
    for Ptrain, Ptest in skf.split(Xtest, ytest):
        print("Fit!\t")
        logistic.fit(Xtest.col[Ptrain, :], ytest.col[Ptest])
        temp.append(logistic.score(Xtest.col[Ptrain, :], ytest.col[Ptest]))
    print("Append!\n")
    accs.append(temp)

accs = np.array(accs)
```



```
Name: Sales, Length: 236380, dtype: int64

In [37]: score = cvs(logistic, Xtrain, ytrain, cv = kf)
D:\Users\sungh\anaconda3\lib\site-packages\sklearn\model_selection\_validation.py:536:
FitFailedWarning: Estimator fit failed. The score on this train-test partition for these
parameters will be set to nan. Details:
MemoryError: Unable to allocate 75.8 GiB for an array with shape (520552, 19542) and data
type int64

  FitFailedWarning)
D:\Users\sungh\anaconda3\lib\site-packages\sklearn\model_selection\_validation.py:536:
FitFailedWarning: Estimator fit failed. The score on this train-test partition for these
parameters will be set to nan. Details:
MemoryError: Unable to allocate 75.6 GiB for an array with shape (520553, 19502) and data
type int64

  FitFailedWarning)
D:\Users\sungh\anaconda3\lib\site-packages\sklearn\model_selection\_
FitFailedWarning: Estimator fit failed. The score on this train-test
parameters will be set to nan. Details:
MemoryError: Unable to allocate 75.7 GiB for an array with shape (52
type int64

  FitFailedWarning)
```

```
  File "<ipython-input-38-e85838054e2b>", line 8, in <module>
    logistic.fit(Xtest.col[Ptrain, :], ytest.col[Ptest])

  File "D:\Users\sungh\anaconda3\lib\site-packages\pandas\core\gene
__getattr__
    return object.__getattribute__(self, name)

AttributeError: 'DataFrame' object has no attribute 'col'


In [39]: Xtest.loc[0:236380, :]
Out[39]:
        Customers  Open  DayOfWeek
0             555     1          1
1             625     1          1
2             821     1          1
3            1498     1          1
4             559     1          1
...           ...   ...        ...
236375          0     0          1
236376          0     0          1
236377          0     0          1
236378          0     0          1
236379          0     0          1

[236380 rows x 3 columns]
```

Choose Logistic Regression

KFold, StratifiedKFold

Compare Accuracy

Modify variable sets

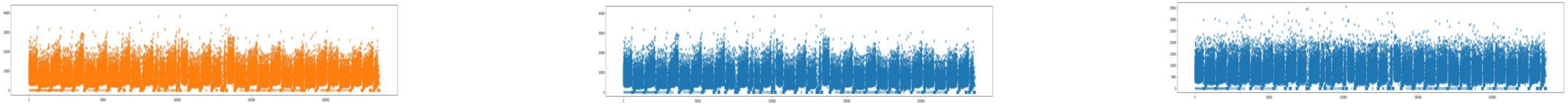
Find better variable sets

Select explanatory variables Select learning methods Summarize

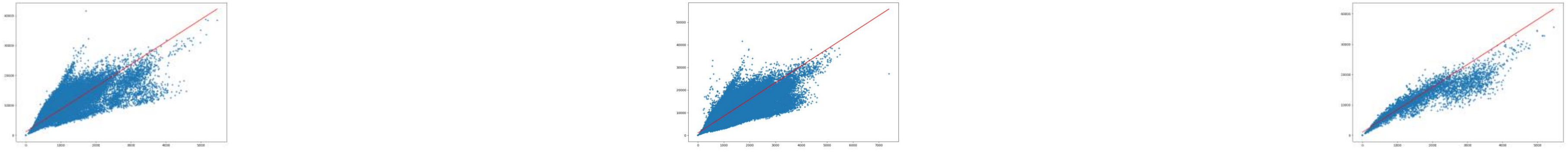
Linear Regression

Actual data

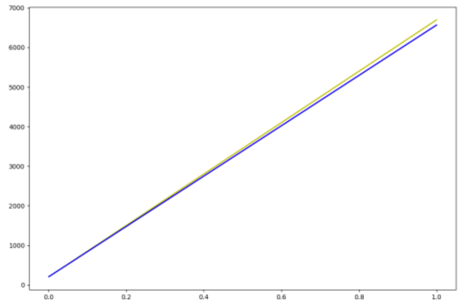
KNeighbors Regression



Scatter plot with regression line for each Sales ~ Dates



Scatter plot with regression line for each Sales ~ Customers



Regression lines for each
Sales ~ DayOfWeek (Binary)

R^2: 0.82970197

R^2: 0.80344976