Engineering New Features

Sreejith Sreekumar

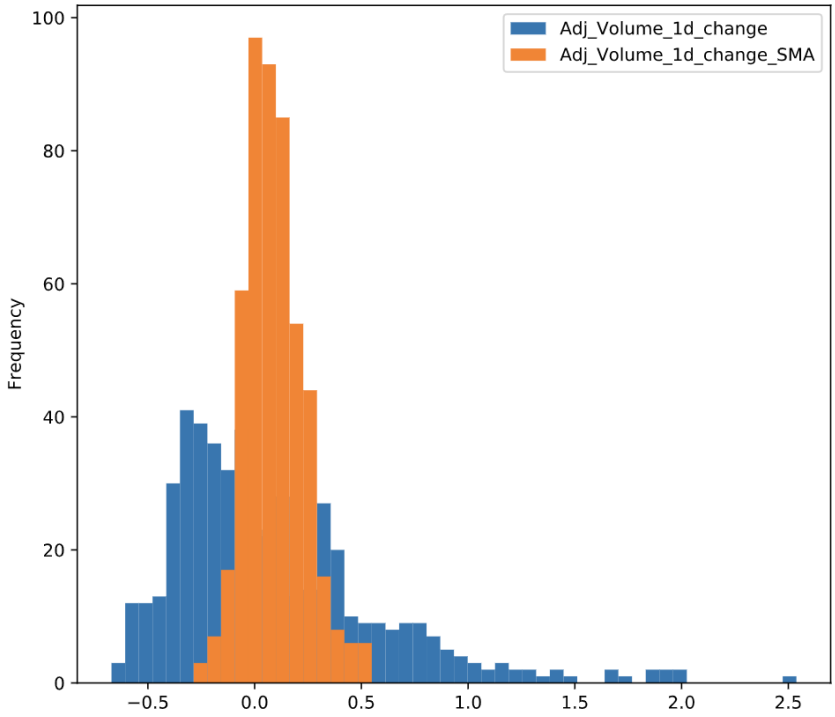
August 13, 2021

Contents

- Linear models cannot include variable interations. This forces us to add variable interations explicitly
- Models like RF, GB Trees, NNs do not require this

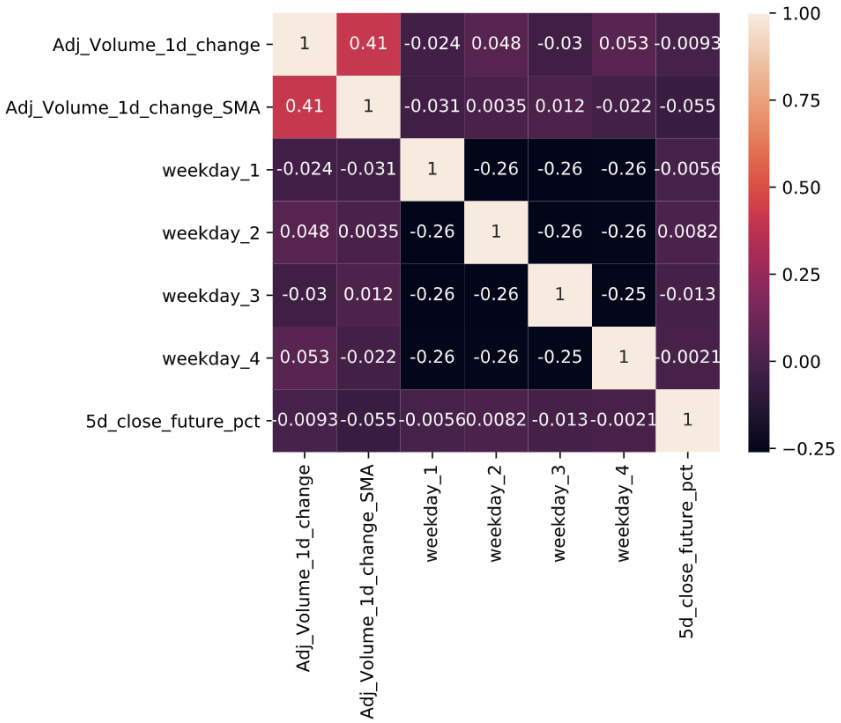
1 Using volume (Number of stocks traded per day) as one of the features

```
# Create 2 new volume features, 1-day % change and 5-day SMA of the % change
new_features = ['Adj_Volume_1d_change', 'Adj_Volume_1d_change_SMA']
feature_names.extend(new_features)
lng_df["Adj_Volume_1d_change"] = lng_df['Adj_Volume'].pct_change()
lng_df["Adj_Volume_1d_change_SMA"] = talib.SMA(lng_df["Adj_Volume_1d_change"].values,
                        timeperiod=5)

# Plot histogram of volume % change data
lng_df[new_features].plot(kind='hist', sharex=False, bins=50)
plt.show()
```



2 Weekdays as Features

```
# Add the weekday labels to the new_features list
new_features.extend(["weekday_" + str(i) for i in range(1, 5)])

# Plot the correlations between the new features and the targets
sns.heatmap(lng_df[new_features + ['5d_close_future_pct']].corr(), annot=True)
plt.yticks(rotation=0)  # ensure y-axis ticklabels are horizontal
plt.xticks(rotation=90)  # ensure x-axis ticklabels are vertical
plt.tight_layout()
plt.show()
```


Adj_Volume_1d_change
Adj_Volume_1d_change_SMA
weekday_1
weekday_2
weekday_3
weekday_4
5d_close_future_pct
1 0.41 -0.024 0.048 -0.03 0.053 -0.0093
0.41 1 -0.031 0.0035 0.012 -0.022 -0.055
-0.024 -0.031 1 -0.26 -0.26 -0.26 -0.0056
0.048 0.0035 -0.26 1 -0.26 -0.26 0.0082
-0.03 0.012 -0.26 -0.26 1 -0.25 -0.013
0.053 -0.022 -0.26 -0.26 -0.25 1 -0.0021
-0.0093 -0.055 -0.0056 0.0082 -0.013 -0.0021 1
1.00
0.75
0.50
0.25
0.00
−0.25