モデル選定について

- 今回はGBDTモデルを作成していこうと思います
    - 選定理由
        - 思ったよりカテゴリ変数が多くOne-HotEncodingしたら次元が大きくなり過ぎる
        - 調整パラメータがそこまで多くない

工夫点

- トレンドを確認してIsHolidayの種類で売上の影響度が異なると感じたので最も影響が大きそうな感謝祭の週かどうかのフラグを追加した
- 欠損値補完の方法をカラム毎で分けることでより妥当性のある補完にしようとした
- Deteに関してコレログラムを作成し周期性を確認した上で特徴量として追加した

予測結果まとめ

- 評価にはMASを使用して予測結果を確認した

```
$ python main.py
n_estimators: 50
Accuracy on training set: 0.743
Accuracy on test set: 0.746
MAE on training set: 6903.394
MAE on test set: 6905.223
n_estimators: 100
Accuracy on training set: 0.743
Accuracy on test set: 0.746
MAE on training set: 6903.394
MAE on test set: 6905.223
n_estimators: 150
Accuracy on training set: 0.743
Accuracy on test set: 0.746
MAE on training set: 6903.394
MAE on test set: 6905.223
n_estimators: 200
Accuracy on training set: 0.743
Accuracy on test set: 0.746
MAE on training set: 6903.394
MAE on test set: 6905.223
n_estimators: 250
Accuracy on training set: 0.743
Accuracy on test set: 0.746
MAE on training set: 6903.394
MAE on test set: 6905.223
max_depth: 3
Accuracy on training set: 0.743
Accuracy on test set: 0.746
MAE on training set: 6903.394
MAE on test set: 6905.223
max_depth: 5
```

```
Accuracy on training set: 0.875
Accuracy on test set: 0.875
MAE on training set: 4626.716
MAE on test set: 4652.046
max_depth: 7
Accuracy on training set: 0.935
Accuracy on test set: 0.931
MAE on training set: 3221.658
MAE on test set: 3274.583
max_depth: 9
Accuracy on training set: 0.967
Accuracy on test set: 0.959
MAE on training set: 2260.030
MAE on test set: 2384.275
max_depth: 11
Accuracy on training set: 0.983
Accuracy on test set: 0.969
MAE on training set: 1588.929
MAE on test set: 1855.710
max_depth: 13
Accuracy on training set: 0.991
Accuracy on test set: 0.972
MAE on training set: 1115.502
MAE on test set: 1622.391
```

課題/モデル考察

- 精度がそこまで良くないので特徴量分析が甘い気がした
    - MerkDownに対す分析を省略したのでここら辺に時間がかけて分析を熱くするべきだった
    - 割と全体的にデータを見てしまっていたので店舗毎、部門毎の分析が甘かった
- 今回はhold-out方でバリデーションを実施したが、本来クロスバリデーションを実施した方が良かった
    - ある程度時系列も関連していそうなのでシャッフルするのも誤りだった
- パラメータチューニングはベイス最適化で実施した方がより効率的にチューニングできたような気がする
    - さらに今回は調整するパラメータを絞って実施したので時間があれば他のパラメータも確認していきたいと感じた
    - max_depthに関してmax_depth > 9にするとtrainデータへの当てはまりが強くなり過学習っぽい感じになっているので他のパラメータで精度を上げるようにした方が良さそう