需要予測
実践チュートリアル

Recruit Restaurant Visitor Forecasting

Index

コンペティション概要

コンペティション概要

・日本の会社リクルートが開催した，レストランの来客数予測コンペ

・ホットペッパーグルメ(hpg)と，Airレジ(air)というリクルートが提供するPOS　レジアプリのデータを用いて予測を行う

・評価指標はRoot Mean Squared Logarithmic Error(RMSLE)

評価指標

・Root Mean Squared Logarithmic Error(RMSLE)という指標で評価を行う

・一般によく用いられるRMSEと比較して以下の特徴がある

・実際の客数より少なく予測した場合，より大きなペナルティを与える

->予測を外すことで仕入れや人員が不足する事態は避けたい

・客数の分布にかなり偏りがあるので，目的変数の分布を正規分布に近づけたい意図もあるかも

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(p_i+1) - log(a_i+1))^2}$$

データ概要

・レストラン，日付ごとのMulti-Index

・hpgとairの予約情報

  ・予約客数
  ・予約した時間
  ・予約が行われた時間
  ・各予約での来客数

・各レストランの市町村情報，緯度経度，ジャンル(和食，フレンチなど)

・曜日や祝日を表すデータ

・実際の来客データ(これを予測する)

EDA

EDAとは

・Exploratory Data Analysisの略で，探索的データ分析と訳される

・データを分析し，データと現象の関係を見出すこと

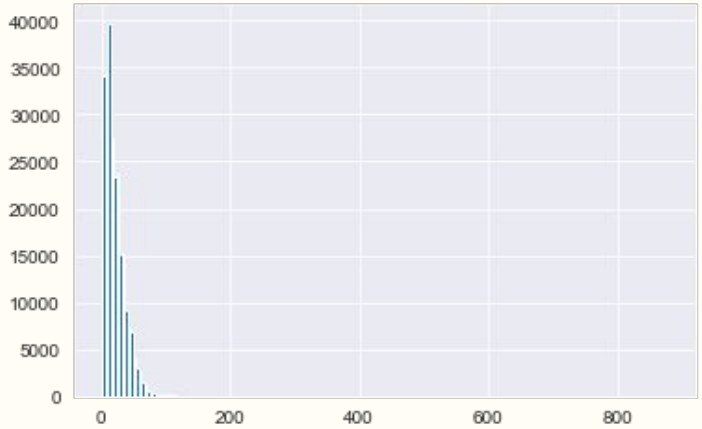
　・データから現象の理解を進め，ビジネスに適用

　・機械学習モデルを構築する際のヒントに　<-　今回は主にこっち

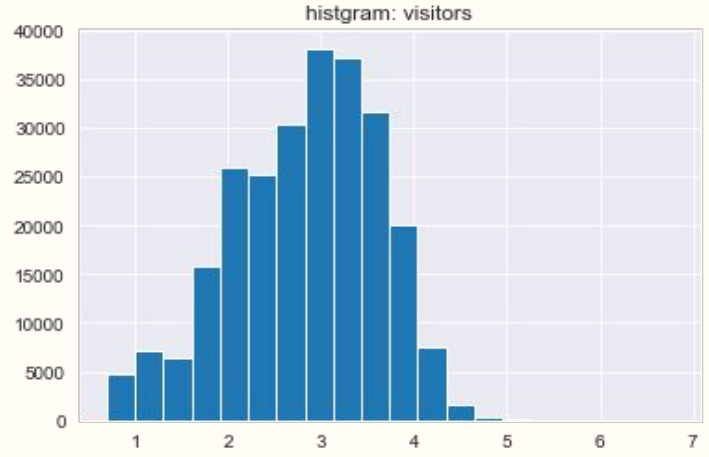
目的変数の分布

・来客数平均は20人程度だが，100人を超える来客もかなり多い

・対数変換log(y+1)を行うと，右図のような綺麗な分布に



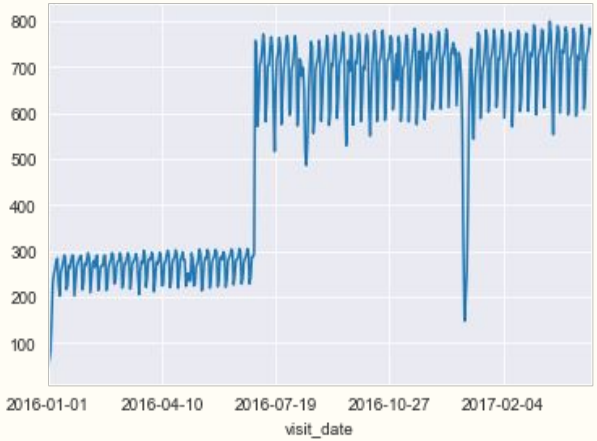
対数変換

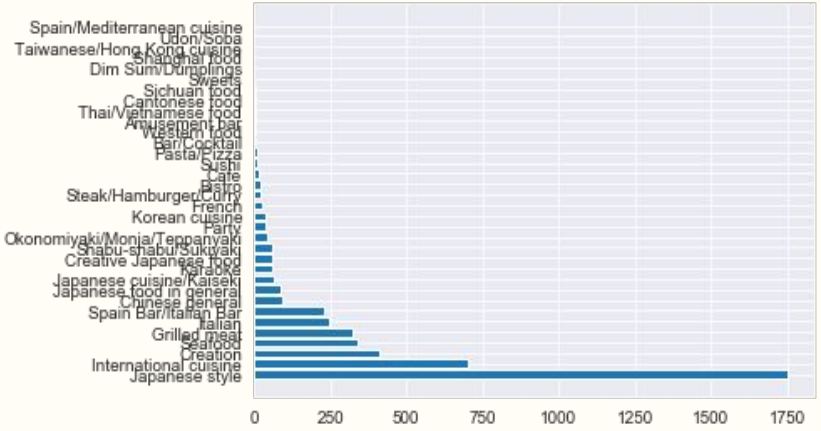


レストランの情報

・日付ごとのレストランの数が，2016年6月ごろに激増している

・hpgはairよりも詳細なジャンルを記述しており，出現頻度にかなりの差がある





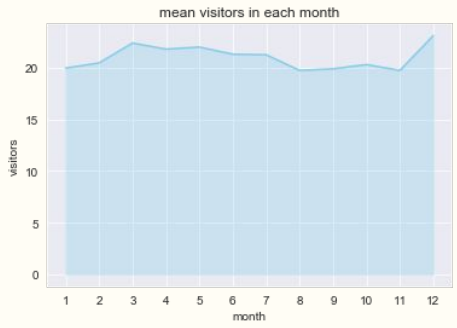
日付情報

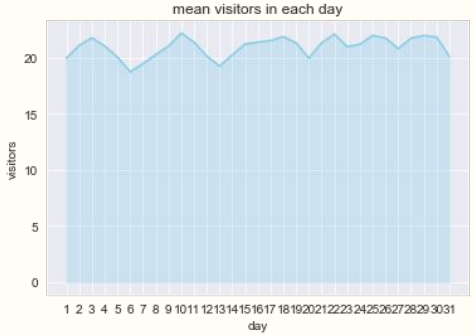
・3月や12月の客数が多い(忘年会や送別会など？)

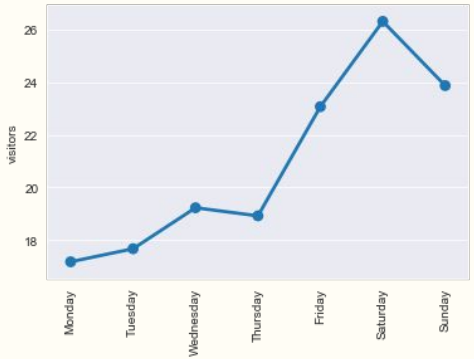
・月のはじめよりも，月の後半の方が客数が多い

・曜日ごとでは金土日がやはり客数が増える

・平日でも特に火曜日が休みの場合に客数が増加する







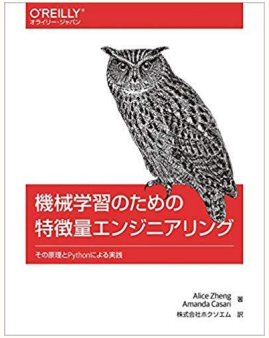
データ前処理と特徴量エンジニアリング

推薦書籍

・データ操作のだいたい全てを網羅した本

・テーブルのデータ操作，データ型ごとの処理の仕方が記載されている

・実践しながら学んでいけばいいが，体系的にまとまっているので目を通してて　損はない



・特徴量エンジニアリングの基本的な手法が網羅されている本

・テキストデータに関する記載もある

・実務やKaggleに取り組んでいくとこの本の言ってる意味がわかるはず

特徴量エンジニアリング

・緯度経度が平均からどれだけ離れているか

・日付特徴量をsin, cosでエンコーディング

・祝日情報と土日の休日情報，翌日が休みか否か，前日が休みか否か

・hpgとairのジャンルを結合した，さらに詳細なジャンル

・各レストランの直近n日の平均来客数(移動平均)

・カテゴリカル変数はCountEncodingとLabelEncoding