Restaurant Visitor Forecasting
(IU Dining Motivated Project)

Shriya Mandarapu, Karthik Shathiri
Luddy School of Informatics, Computing, and Engineering
Department of Data Science
Indiana University Bloomington
{shmanda, kashat}@iu.edu

Dr. Yuzhen Ye
Luddy School of Informatics, Computing, and Engineering
Department of Computer Science
Indiana University Bloomington
yye@iu.edu

***Abstract*—This report analyses various approaches to dealing with time series data. Long historical data can be run with either Regression models or considered as a time series problem. After studying classic approaches such as ARIMA, SARIMA and their stationarity testing criteria we moved onto advanced models such as Prophet. Prophet is run with multiple choice of parameters and additional regressors and the performance is evaluated. We also built multiple Regression models such as XGBoost and AdaBoost and compared the results across different models.**

***Index Terms*—Prophet, Forecast, GAM, Restaurant Visitor Prediction, Kaggle, Regression, RMSLE, XGBoost, AdaBoost, Cross Validation, Grid Search CV**

I. MOTIVATION

Having worked at IU Dining for about 3 months now and watching every operation closely, we have observed and noted several kinks and inefficiencies in the process that we believe could be solved by taking a data driven approach. Motivated to apply theoretical knowledge to solving a real life problem, we approached the IU Dining Administration Director Mr. Rahul Shrivastav who was very much thrilled with the idea.

Our proposed objective was to solve 5 major current issues at IU Dining -

- Food Wastage Reduction
- Effective Utilization of Workforce
- Meal Plan Engineering
- Inventory Management
- Improve Quality of Service

Demand forecasting and analyzing food consumption patterns and Inventory loss across eateries were the initial ideas.

Although the project did receive approval and we were able to get our hands on some overall revenue figures. POS data which is a crucial piece is sensitive information and IU policies require any user to undergo data security training before accessing the database. These additional permissions and policies pushed the timeline of the project execution to the end of November, forcing us to pick a publicly available data set that is most representative of our use case and analysing it to test various approaches. Nevertheless, the IU Dining project which started as a course project will now continue to be our priority and we hope to revamp the current workings at IU Dining over the course of time, giving it a new data driven face.

II. INTRODUCTION

Keeping our objective at the forefront, we chose a dataset for restaurant visitor forecasting spanning over a period of one year. While there are clear differences in the underlying working assumptions of IU Dining to that of local restaurants, we believed it would still be a sufficiently close case study dealing with similar marked conditional seasonality trends. We picked the “Recruit Restaurant Visitor Forecasting” dataset from Kaggle published in 2017. We tried multiple existing approaches and new models to evaluate which gave the best performance and why.

A. WHAT HAVE WE DONE DIFFERENTLY FROM OTHER KAGGLERS?

The data set that we have considered was published in 2016. Most Kagglers have used existing traditional time series models such as ARIMA and SARIMA for their analysis or regression models for prediction. Prophet is an additive model based, forecasting time series model, first published in 2018 by researchers at Facebook. What’s interesting is that prophet considers the data as f(t) instead of the regular lag dependencies we account for in most time-series approaches. Which thereby makes it robust at handling missing data and outliers. It does allow inclusions through additional regressors to capture other predictor influences such as holidays, weather, etc. We tried to observe what insights this unique approach to tackling time-series will actually provide.

III. Brief Literature Review

A. Classic Approaches - ARIMA & SARIMA

Auto Regressive models such ARIMA and SARIMA are the go to techniques for most time series forecasting problems. However they are not without drawbacks. Most notable of them include sensitivity to outliers and missing data, which is very much the problem in our data set. The data also show a strong seasonality. ARIMA does not support seasonal data. Though SARIMA was designed to handle seasonality, one can't include additional features that affect the data eg. holiday effects or weather etc. Hence to circumvent these issues we choose prophet as our forecasting tool.

B. Prophet

Prophet decomposes the time series into nonlinear additive trends using partial fourier sums that fit with daily, weekly and yearly seasonality, plus holiday effects and additional regressors if integrated with the input model. Conditional seasonality is one additional feature that is brilliant about it - It makes it possible to study seasonality trends for two different conditions, say to observe dining visitor trends during examination weeks vs non-examination weeks. This feature to custom seasonalities by adding additional regressors is where domain expertise can help build a rather intelligent model rather than plugging the entire universe of parameters. It has additional parameters to adjust the regressor impact to control over and underfitting such as holiday_prior_scale and changepoint_prior_scale.

IV. Data Set Summary

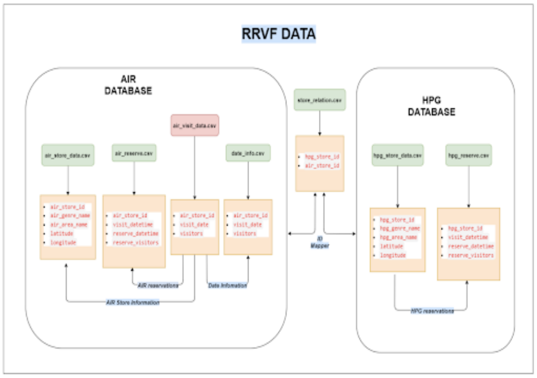


Fig. 1. Restaurant Visitor Data

We have data from two reservation systems, Hot Pepper Gourmet (hpg): here users can search restaurants and also make a reservation online and AirREGI / Restaurant Board (air): a reservation control and cash register system.
To briefly outline the Data Set in our analysis -

- Training data period - 2016-01-01 to 2017-04-22
- Test data period - 2017-04-23 to 2017-05-31
- Test data spans golden week (End of April & Early May)
- Reservations data available for both hpg & air
- Visitors data available for air system
- Tables to merge both system information (IDs)
- No missing values but missing dates
- Holiday flag data

Train Data set

- Total number of unique AIR restaurants - 829
- Total unique genre in AIR restaurants - 14

Test Data set

- Total unique restaurants:- 821

V. Exploratory Data Analysis

Breaking it down into four major pieces of study -

- Understanding the hotel base
- Underlying trends within and in-between the reservation systems
- Customer Behavior + Weekend Effects
- Genre Impact
- Holiday Effect

A. Understanding Hotel Base

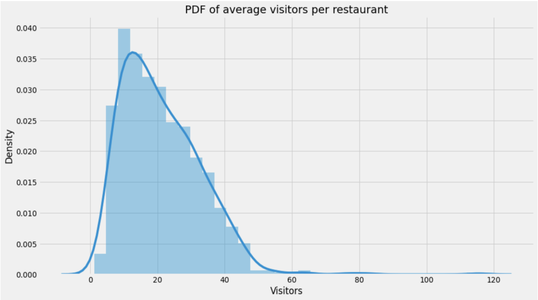


Fig. 2. Hotel Visitor PDF

Decently normal distribution with slight right skewness and a mean visitor count around 20. Almost 90% of the restaurants have less than 40 visitors/day. This is indicative of the fact that most hotels in our data set are small businesses.

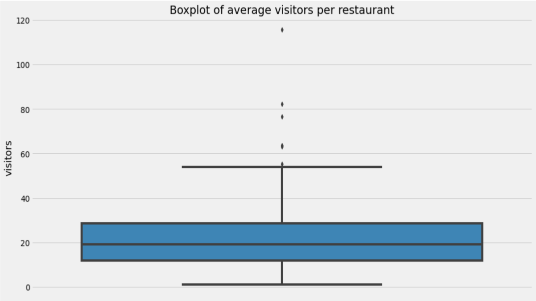


Fig. 3. Hotel Visitor Boxplot

B. Within and in-between Reservation Systems

Studying the differences in the reservation patterns between the two systems:

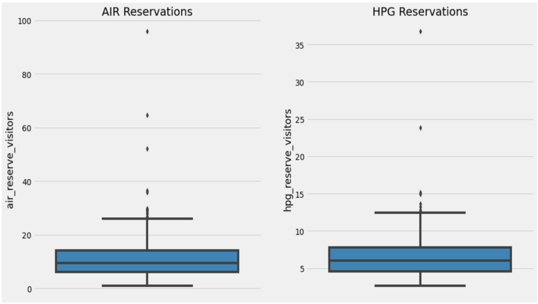


Fig. 4. Air vs Hpg

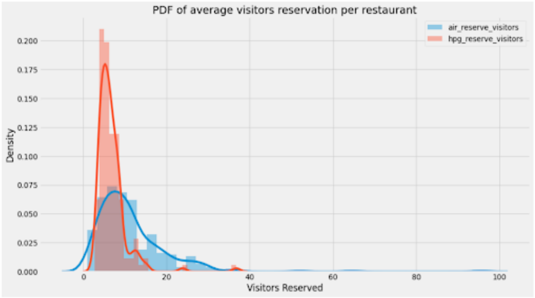


Fig. 5. PDF Plot

C. Customer Behavior + Weekend & Holiday Effects

Saturday sees the highest visitors. Sunday is surprising low. There is certain amount of variation through out the year. December appears to be the most popular month for restaurant visits, this could be because December in general is a festive month. The period of Mar - May is consistently busy. We also observe a marginal increase towards the 25th of the month, which is when salary gets credited in Japan. But there is no significant holiday effect. This could probably be attributed to the Japanese value systems of spending the holidays with family rather than stepping out. There is a clear weekend effect though, so if there is a holiday that falls on the weekend, the demand variation can be attributed more to it falling on the weekend than for being a holiday.

Some additional observations observed in the reservation patterns shows a nice 24 hour gap between reservation and visit for dinner reservation on both systems and mostly these reservations come in the evening hours 24 hours in advance.

D. Cuisine Impact

There are 14 different cuisines with Izakaya being the most popular one making up almost 23.8% of the base. Interesting observation on cuisine choices over weekdays and weekends show clear spike in demand for Karaoke bars and International cuisines over the weekends.

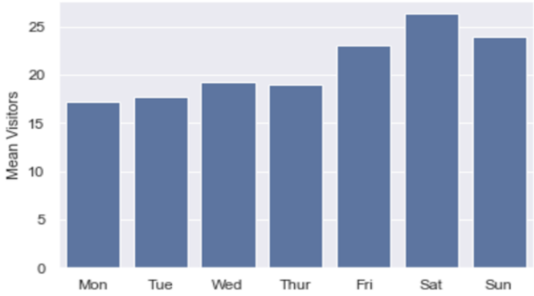


Fig. 6. Day of Week

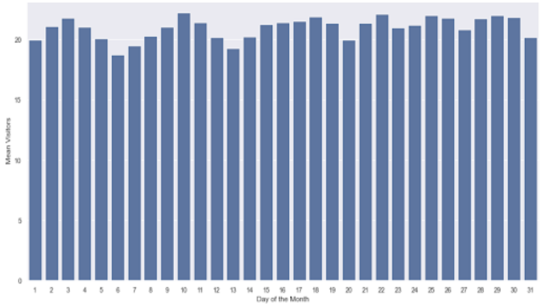


Fig. 7. Day of Month

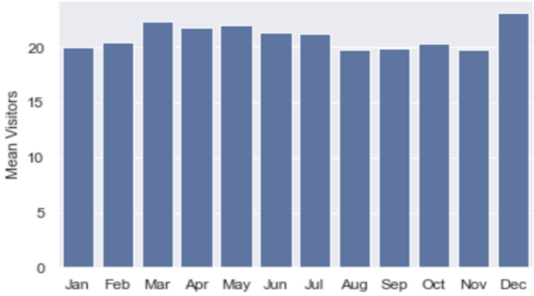


Fig. 8. Monthly Trend

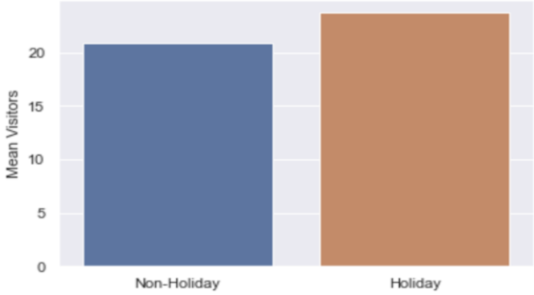


Fig. 9. Holiday vs Non-Holiday

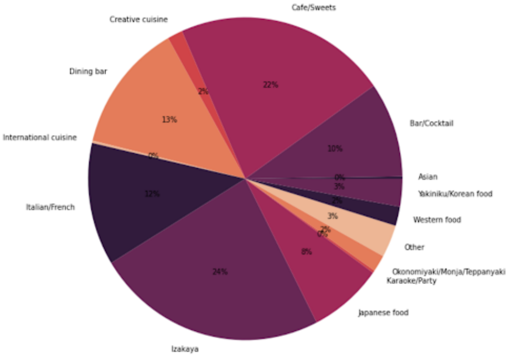


Fig. 10. Cuisine Chart

E. Holiday Effect

It can observed from this above plot that pre post holidays are good markers for holiday demand. There is a slight spike in visitor count just one day prior to a holiday in most cases.

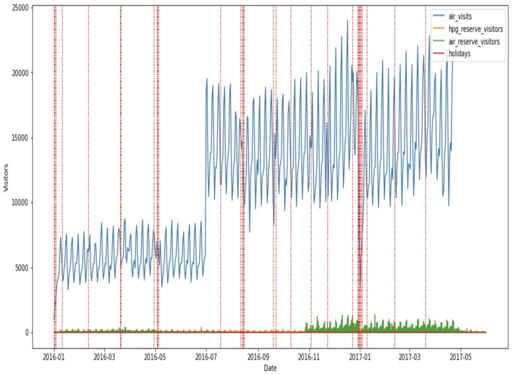
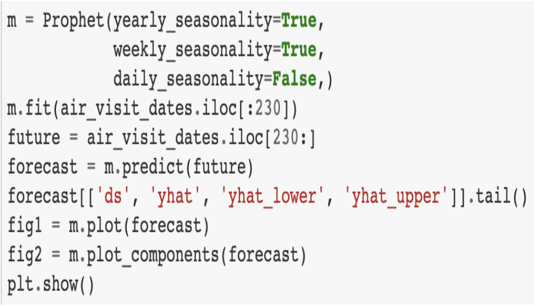


Fig. 11. Pre-Post Holiday Markers

VI. Time Series Approach

We ran prophet using several parameters. Prophet takes the data input as a dataframe of two columns with specific labels "ds" and "y", ds being the date and y being the prediction variable, here in this case we took the air visitor information.



```
m = Prophet(yearly_seasonality=True,
            weekly_seasonality=True,
            daily_seasonality=False,)
m.fit(air_visit_dates.iloc[:230])
future = air_visit_dates.iloc[230:]
forecast = m.predict(future)
forecast[['ds', 'yhat', 'yhat_lower', 'yhat_upper']].tail()
fig1 = m.plot(forecast)
fig2 = m.plot_components(forecast)
plt.show()
```

Fig. 12. Prophet model Syntax

Starting with the basic prophet model taking no additional regressors the model performance was evaluated. We later including the holidays data as an additional regressor, which showed improved performance. To further enhance the prediction, we scraped weather data for the same time period and integrated the values of temperature and precipitation and ran them as added regressors.

VII. Regression Approach

A. Prophet to Regression Switch?

Prophet fails to take into account additional dependencies that could have significant impact on the visitor forecast. Eg: addition of a competitor restaurant in the same area could decrease the visitors for a restaurant or the genre of restaurant could explain its seasonality eg. cafe/sweet type restaurants could see an uptick in the summer. To include such external features this forecasting can be thought of as a regression problem.

B. Brief Introduction

Ensemble methods using decision trees is a popular way to solve regression problems. The method combines a set of weak learners into a strong learner to minimize training errors. We explore boosting methods in this project, AdaBoost and XgBoost in particular.

In boosting, after each iteration of training each model tries to compensate for the weaknesses of its predecessor. With each iteration, the weak rules are combined to form one, strong prediction rule.

C. Feature Engineering

Now taking a regression based approach requires us to include additional features for better prediction. Based on the above exploratory analysis, some do particularly stand out and look promising. We have included features such as

- Hour gap between reservation time & actual visit time
- Day of week, Month, Year break up from date
- Pre-holiday Post-holiday flags
- Visitor summary statistics
- Weather Data - Temperature
- Weather Data - Precipitation
- Distance from weather station
- Day of month
- Categorical One hot encodings

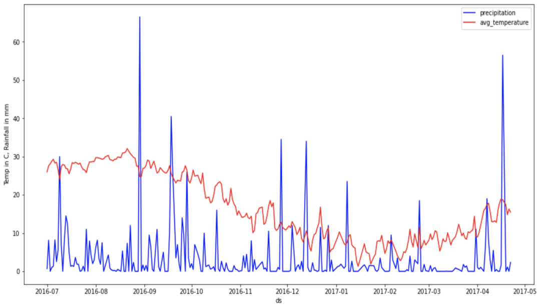


Fig. 13. Weather Data

VIII. Work Done and Results Analysis

A. Prophet

Prophet - Considering only weekly seasonality trend keeping yearly and daily seasonalities turned off & No additional regressors :

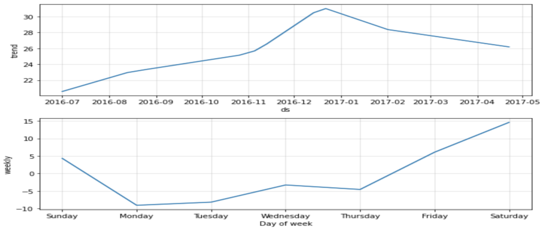


Fig. 14. Prophet - Only Weekly Regressor

Prophet model taking flat growth rate and in-built holiday regressor -

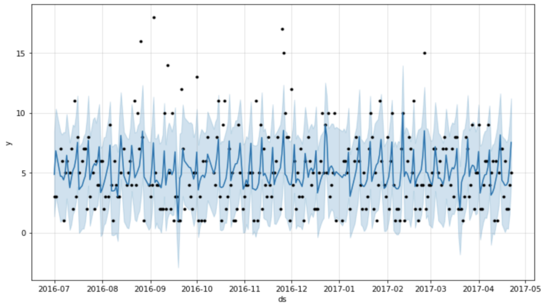


Fig. 15. Prophet Model - In-built Holiday - Flat Growth Forecast

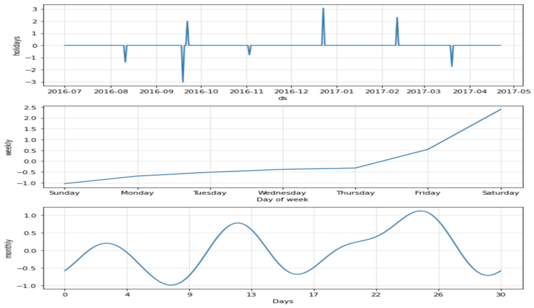


Fig. 16. Prophet components - Flat Growth

B. Regression

We tuned the hyper parameters for these models using Grid search and found the following parameter to give the best results.

XGBoost Parameters: learning_rate : 0.01, min_child_weight : 0.8, subsample : 0.6, colsample_bytree : 0.8, max_depth : 8

AdaBoost Parameters: learning_rate : 0.01, n_estimators : 100

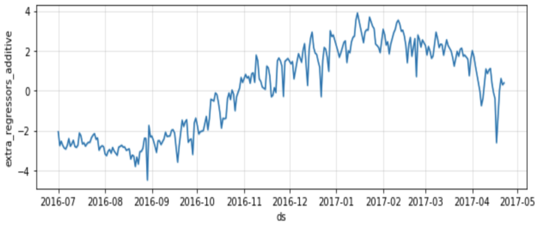


Fig. 17. Prophet - Weather Regressor

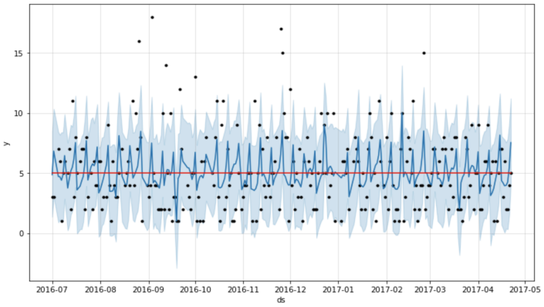


Fig. 18. Prophet Forecast - Flat Growth

To evaluate the models we used a 3 fold cross validation strategy. The predictions for the three models were average for the submission. We observe that Xgboost performs better than adaboost.

C. Prophet Observations

Prophet is an amazing model when the data is atleast a year long. In the dataset that we have, we see a huge number of onboarded properties, about 500, mid 2016. These hotels seem to have poor prediction compared to the ones where we have a larger timeframe of data available. Changepoint priors have been tested. Making it too flexible only worsened the performance.

There were several parameters that we have tried to experiment with, such as, added weather regressors, in-built vs external holiday flag regressor, combinations of yearly, daily, and weekly seasonalities. Given our data is barely a year long in most cases, turning it off worked best.

While trying the external holiday flag inclusion as well as in-built holiday regressor specified with the country index, In-built regressor did a better job at capturing the holiday variations.

We also tried to capture the trend across the dataset in our initial analysis, while assuming a almost flat or zero growth trend over the period, it seemed to capture the subtle crests and troughs. It probably would make more sense to capture the growth trend when the data spanned a wider timeframe with clear trend change.

IX. Results & Conclusions

- We have investigated two approaches to forecast restaurant visitors - Time series and regression based. Both have their own strengths and weaknesses.

- Prophet - works well out of the box, in terms of explainability and is rather intuitive to understand.
- Regression - works better than prophet, require heavy feature engineering that don't necessary make intuitive sense, requires a bit of probing to understand the built model.
- The test data set contains "golden week" - a set of holidays that are observed around late april and early may. The bad performance on the private data set can be attributed to the fact that the training data doesn't include this week for most of the restaurants.

X. MODEL PERFORMANCE

There are several metrics that can be used for model evaluation such as R squared error, Mean Squared Error, Root Mean Squared Error and Root Mean Squared Log Error.

$$RMSLE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(y_i + 1) - \log(\hat{y}_i + 1))^2}$$

Fig. 19. Root Mean Squared Log Error

RMSLE penalizes underpredictions more than overprediction. Useful in cases where underprediction is not acceptable. Therefore, RMSLE will be our choice of evaluation metric since for hotels it's better to be overprepared than underprepared which will directly impact the quality of service.

XI. ACKNOWLEDGMENT

We would like to thank Professor Yuzhen Ye for the constant motivation and flexibility to try and go beyond just class material. We probably wouldn't have taken up this wonderful initiative otherwise. It was an amazing learning experience and journey overall. We hope the dining project will change the course of dining operations and prove to be revolutionary, which we owe entirely to the professor.

References

[1] Source of Dataset : https://www.kaggle.com/c/recruit-restaurant-visitor-forecasting/data

[2] Auckland Cycling - Prophet with weather regressor : https://nbviewer.org/github/nicolasfauchereau/Auckland_Cycling/blob/master/notebooks/Auckland_cycling_and_weather.ipynb

[3] Reference Blog1 : https://csanjay.medium.com/recruit-restaurant-visitor-forecasting-a704cd5432c8

[4] Reference Blog2 : https://medium.com/analytics-vidhya/recruit-restaurant-visitor-forecasting-f9ef87ba1073

[5] Reference Blog3 : https://machinelearningmastery.com/time-series-forecasting-with-prophet-in-python/

[6] Reference Approaches 1 : https://www.kaggle.com/gvyshnya/gbm-cv-submission-lb-score-0-511

[7] Reference Approaches 2 : https://github.com/MaxHalford/kaggle-recruit-restaurant/blob/master/Solution.ipynb

[8] Reference Approaches 3 : https://www.kaggle.com/irinaabdullaeva/welcome-recruit-restaurant-visitor-forecasting/notebook

[9] Prophet Publication : https://peerj.com/preprints/3190/

[10] IU Dining Article - Supply Staff Shortage: https://www.heraldtimesonline.com/story/news/local/2021/09/29/iu-bloomingtons-dining-halls-struggle-supply-staff-shortages/8417866002/

[11] Dining - Mismanagement & Quality Issues: https://www.idsnews.com/article/2021/10/iu-dining-mismanagement-food-quality-issues-empty-promises

[12] Data Driven Dining - Motivation: https://www.mastercardservices.com/en/reports-insights/data-driven-dining-top-trends-restaurants-2019

[13] Weather data Source: https://www.kaggle.com/huntermcgushion/rrv-weather-data?select=weather_stations.csv

TABLE I

MODEL PERFORMANCE

Model	Parameters	Private Score	Public Score
Prophet	Default(Trend change point = 0.05), No added Regressors	0.60180	0.54495
Prophet	Default(Trend change point = 0.05), Holiday Regressor	0.60041	0.54253
Prophet	Trend Changepoint = 0.5 , Holiday Regressor	0.60081	0.53312
Prophet	Trend Changepoint = 0.8 , Holiday Regressor	0.61057	0.53598
Prophet	Trend Changepoint = 0.5, Weekly & Monthly seasonalities , In-built Holidays & Flat growth rate	0.606	0.561
AdaBoost	learning_rate = 0.01, n_estimators= 100	0.55	0.517
XGBoost	learning_rate = 0.01, min_child_weight = 0.8, subsample = 0.6, colsample_bytree = 0.8	0.538	0.501