Expedia Hotel Room rate Prediction

(Final Project of SI699: Big Data Analytics)

Jing-Crystal Wang, Yuan Li
School of Information, University of Michigan

PROJECT OBJECTIVE

The overall goal of this project is to build a hotel room rate prediction system that helps customers to evaluate the price and determine the best time to book a room for traveling. Several questions that we would like to answer include:

- ☐ When we book hotel, how can we know the price is reasonable or overcharged? Can we have a number to benchmark?
- ☐ Can we know the fluctuation of hotel room price by season?
- ☐ ...

DATA OVERVIEW

We use Personalize Expedia hotel searches – ICDM 2013 from online Kaggle competition (> 4GB), which includes a wide variety of data on User, Property, time, competitor, etc.

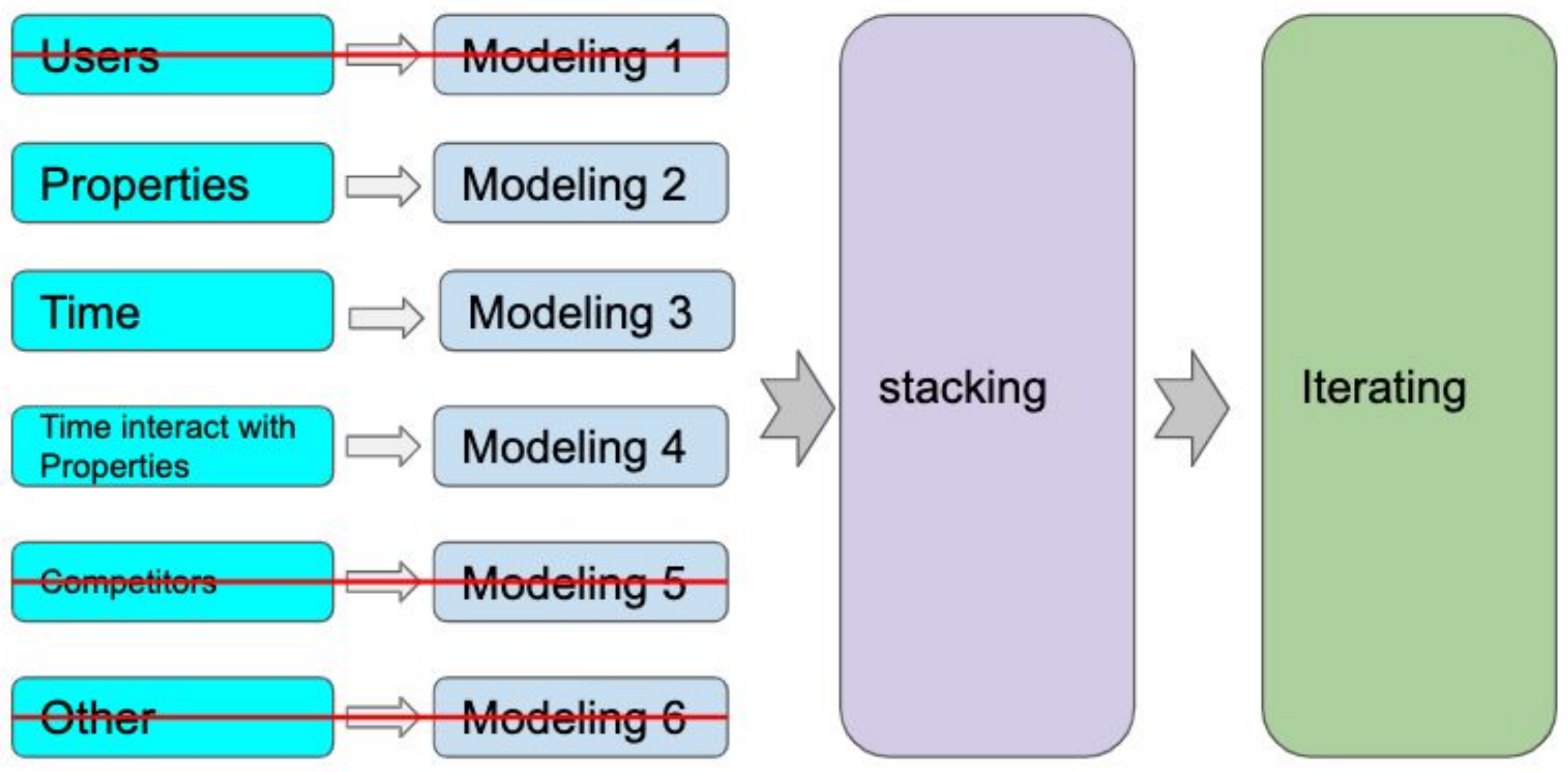
METHODOLOGY

Data Preprocess:

- handle missing data
- Outlier value detection
- Convert categorical variables to continuous variables
- Aggregate data based on time range

Modeling:

modeling for each feature group and stack each model as second-layer model



~~Users~~ ⇨ ~~Modeling 1~~
Properties ⇨ Modeling 2
Time ⇨ Modeling 3
Time interact with Properties ⇨ Modeling 4
~~Competitors~~ ⇨ ~~Modeling 5~~
~~Other~~ ⇨ ~~Modeling 6~~
stacking
Iterating

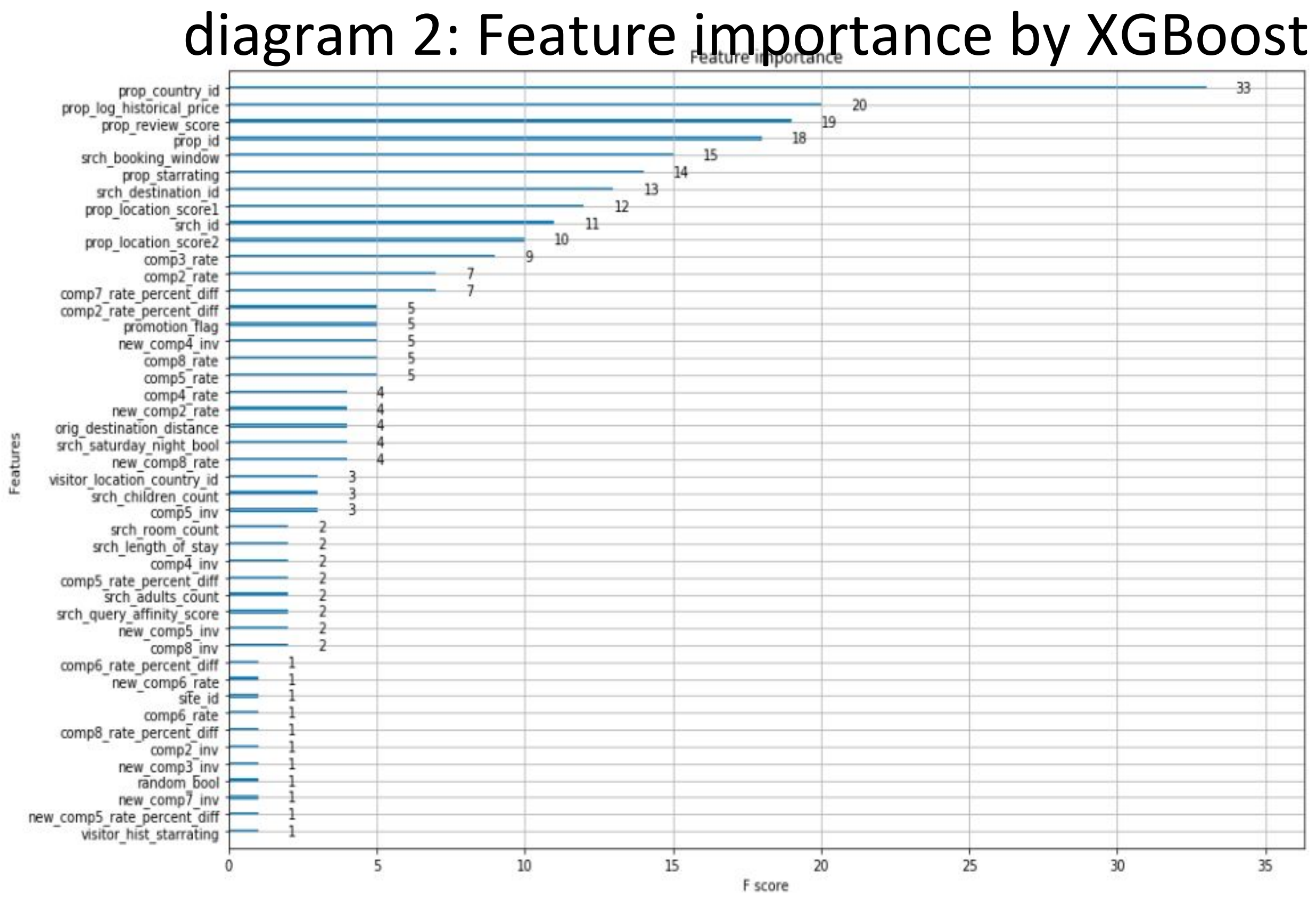
diagram 1: multi-layer modeling structure

RESULTS

table 1: performance of property modeling

	Train RMSE	test RMSE
Ridge	18.80	41.93
Decision tree	24.84	22.50
Random Forest	19.47	22.15
Elastic Net	19.30	17.27

diagram 2: Feature importance by XGBoost



country id, property historical price, property review score are the top 3 most important features

diagram 3: autocorrelation of hotel room rates

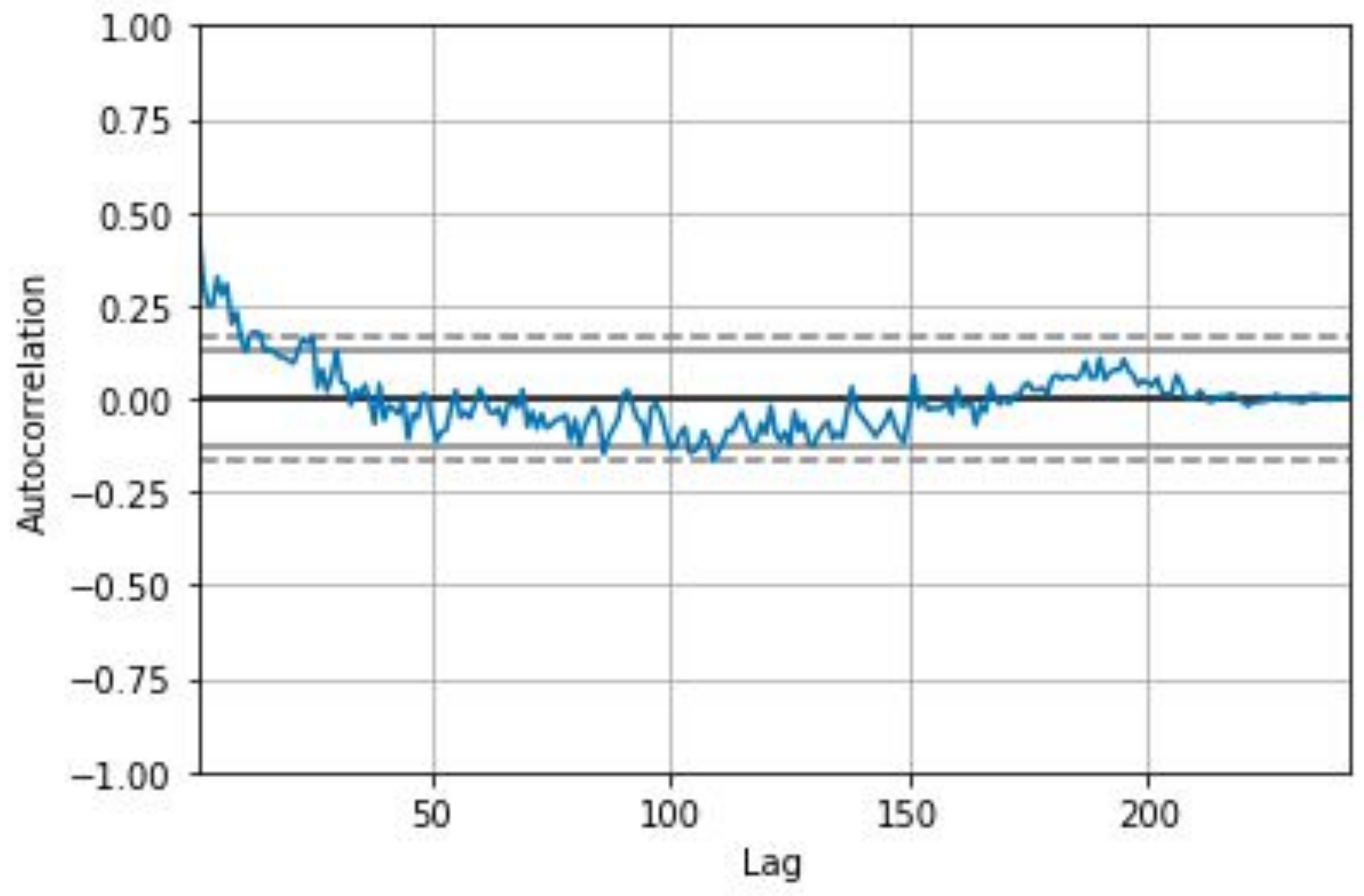
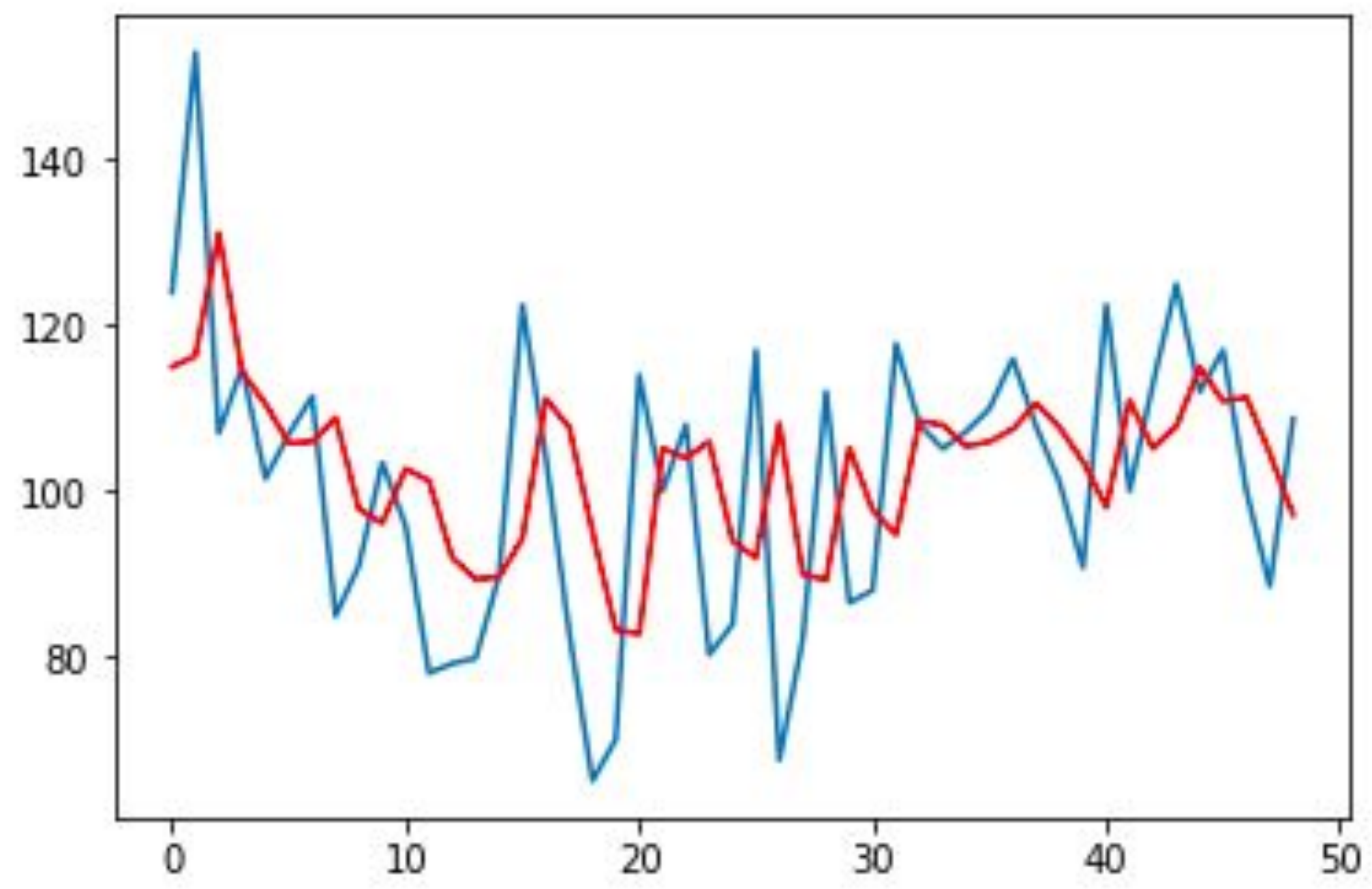


diagram 4: hotel room rate trend and prediction



blue line is the actual room rates and the red line represents the predicted price by ARIMA model

table 2: performance of time modeling

	Train RMSE	Test RMSE
ARIMA	19.734	16.756
Linear Regression	22.87	20.36
XGBoost (as a second-layer modeling)	18	17

table 3: performance of stacked model

	Train RMSE	Test RMSE
linear regression	12	21
Random Forest	5	23
XGboost	22	18

Use predictions from property and time modeling as new features, and feed in a second-layer model

CONCLUSION

For the first-layer modeling, we fit different models based on the characteristics of different feature group. For property feature modeling, Elastic Net shows the best performance, followed by Random Forest. In terms of time feature modeling, we fit ARIMA and linear regression model. Ultimately, we used the result from first-layer models, and fit a second-layer model, which turns out that XGBoost shows the best performance.

REFLECTION

- Hotel price is predictable
- Hotel price is related with hotel's attributes and fluctuation vary by time
- Feature engineering is the key
- Multi-layers modeling to integrate the regression model and time series model

This is a final project presentation of SI 699: Big Data Analytics. We thank Professor Mei for his instruction.

Contacts:
wjingcc@umich.edu
yuanlii@umich.edu