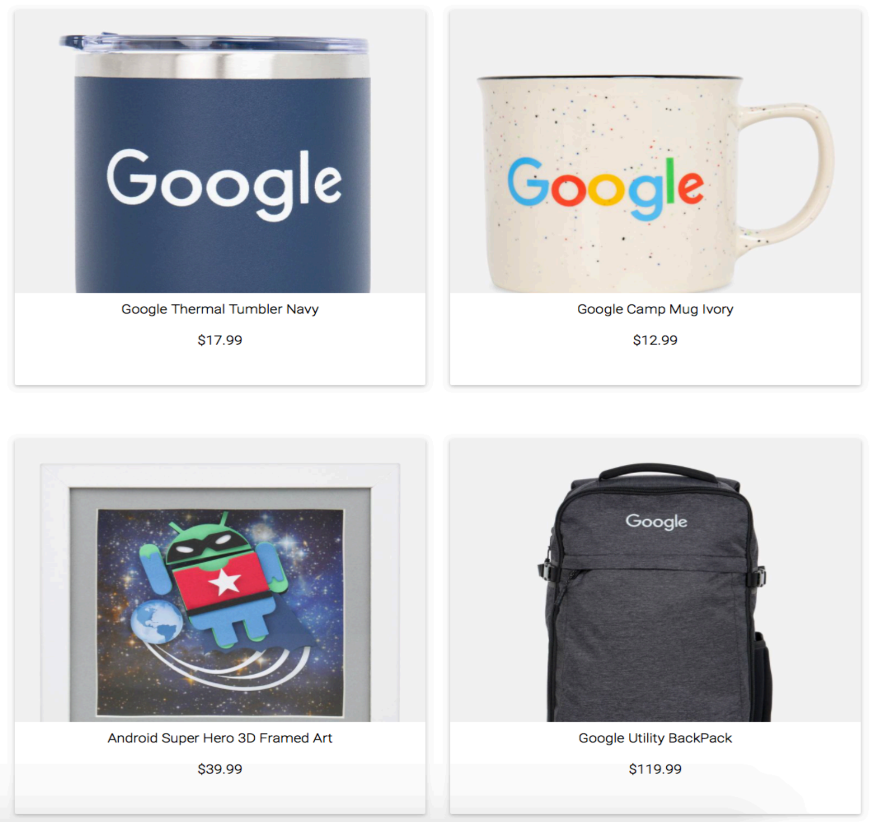
Final Project: Predictive Modeling

Google Analytics Customer Revenue Prediction

Group 24

Zhenru Han, Shiwei Hua, Sitong Liu, Runjie Lyu, Zixiao Wang

Google Merchandise Store

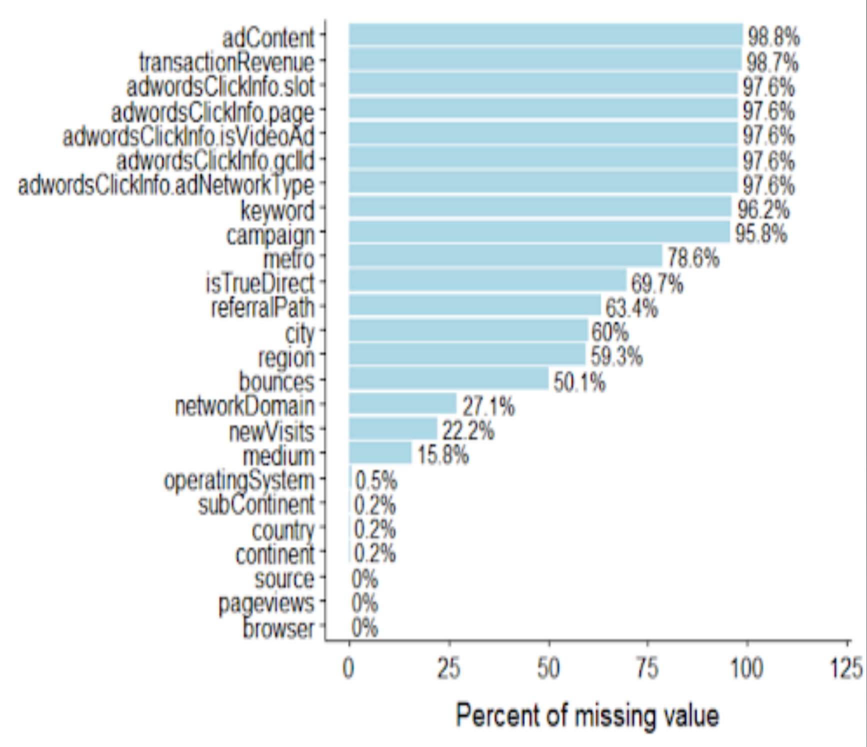


- The dataset is from Google Customer Revenue Competition on Kaggle, and it has been already divided into train and test.
- The train set included 903653 observations with 36 variables and test set contained 804684 observations.

Data Summary

- Contains 6 groups: visitor info, visit number, channel, geo networks, devices, and advertisement.

- Highly Imbalanced Data: 98.7% users did not make the transaction while looking the product.



PROJECT GOAL

Build several statistical and data mining models to evaluate the revenue and transaction of Google Store.

Specific Questions:

1. What would be the daily average revenue in the future 60 days?
2. Whether a customer would make a purchase or not?
3. What would be the purchase amount of each customer?

LOGIC PROCEDURE

Pre-process
(data collection
+ data clean)

24.3 GB → 200 MB

Exploratory Data Analysis

Time Series

Classification Models

- Random Forest
- Logistic Regression
- XGBoost

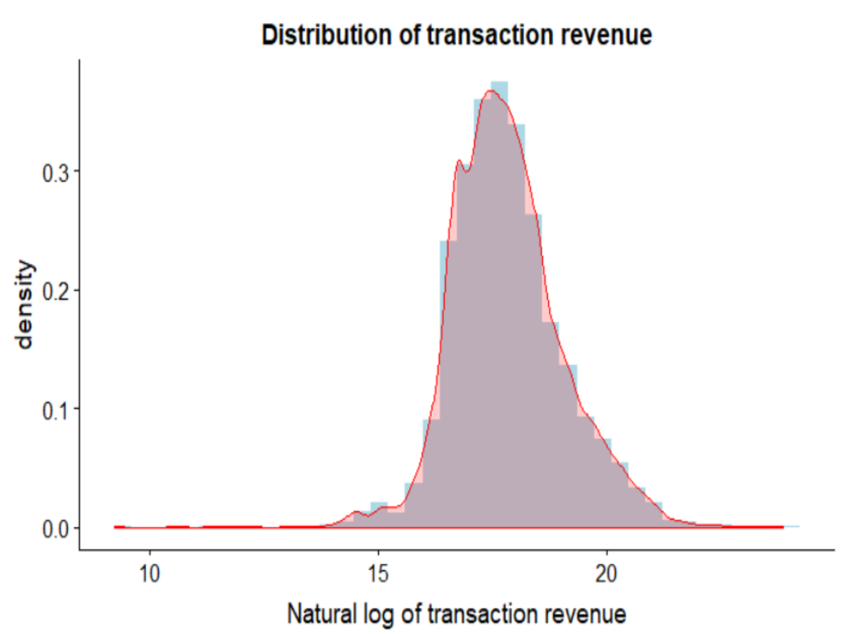
Regression

- LMM
- XGBoost

Conclusion

Discussion

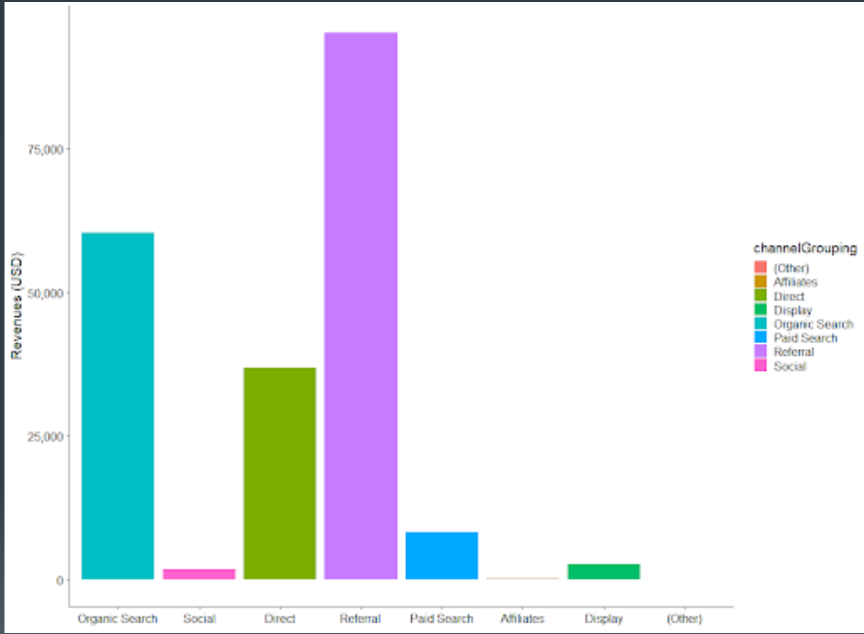
Exploratory Data Analysis

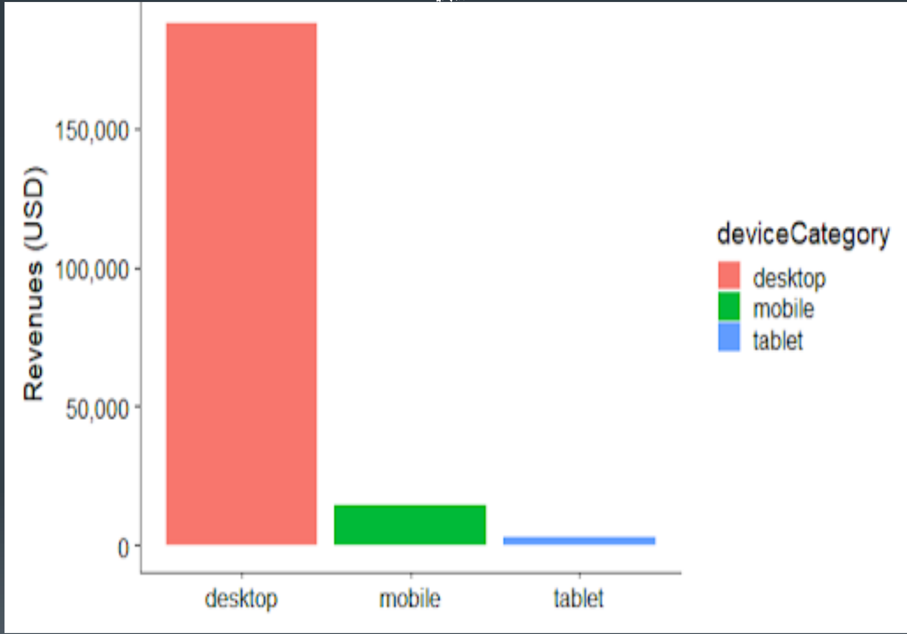


Variable	Transaction Revenue	Log Revenue
Minimum	0.01	9.2
1st Quantile	24.9	17.0
Median	49.5	17.7
Mean	133.7	17.8
3rd Quantile	107.7	18.5
Maximum	23129.5	23.9

- Data transformation: “log1p” function transforms x to log(1+x).

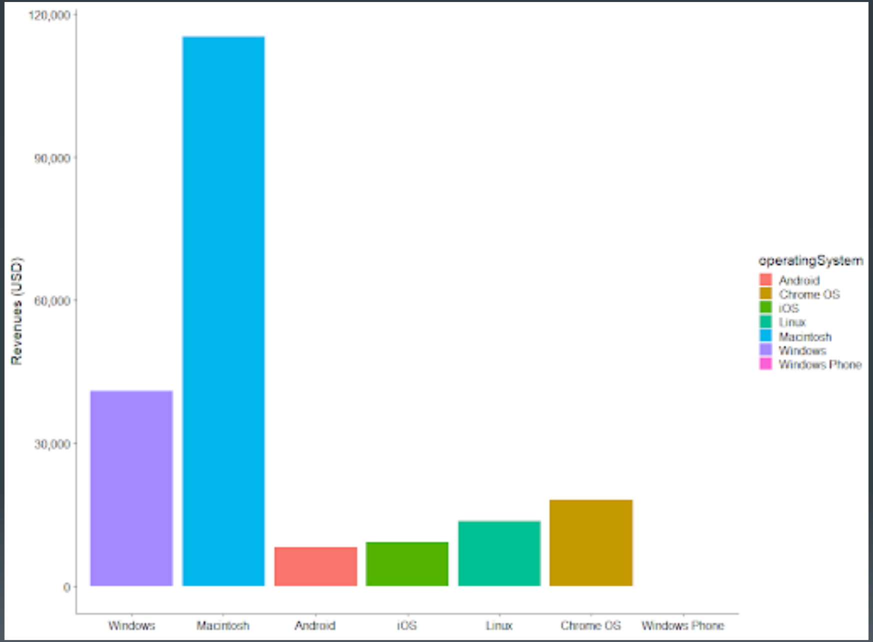
Exploratory Data Analysis

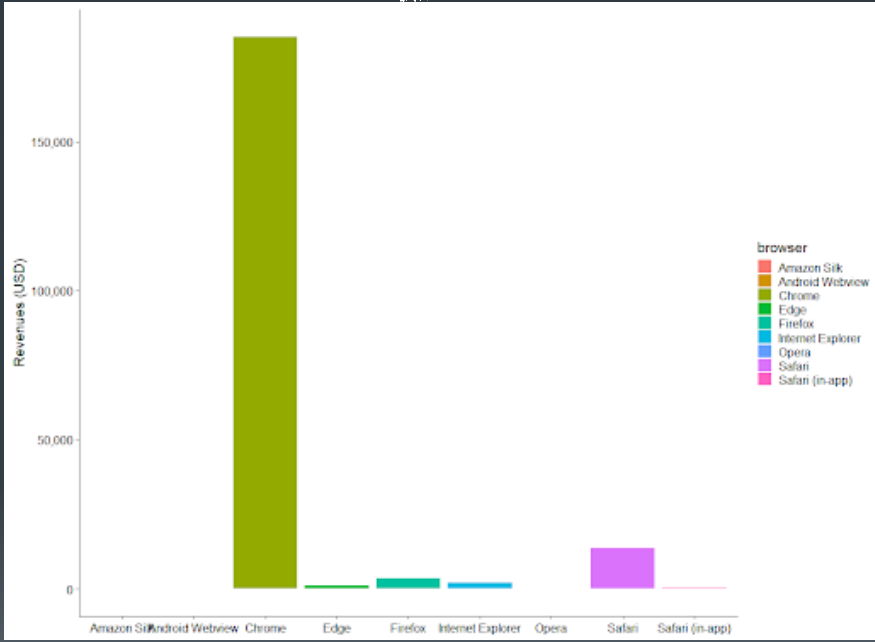




- Most number of sessions come from “organic search”, but “referral” contributes most transaction revenue.

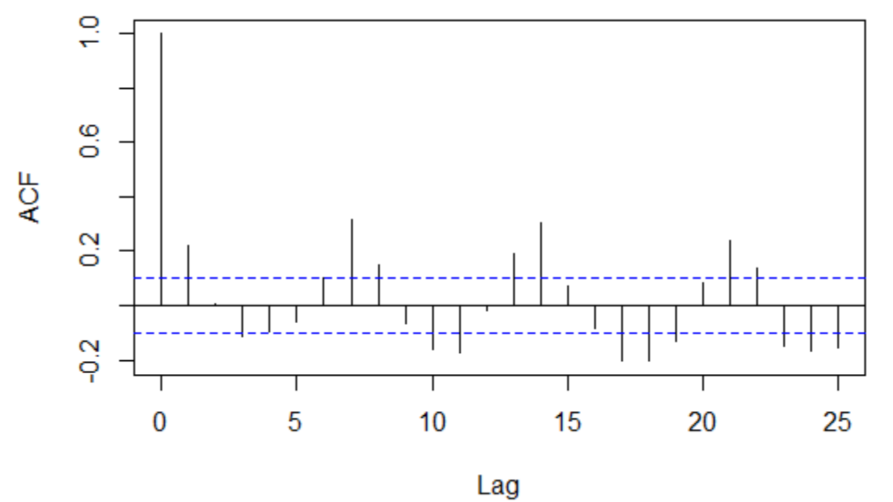
Exploratory Data Analysis

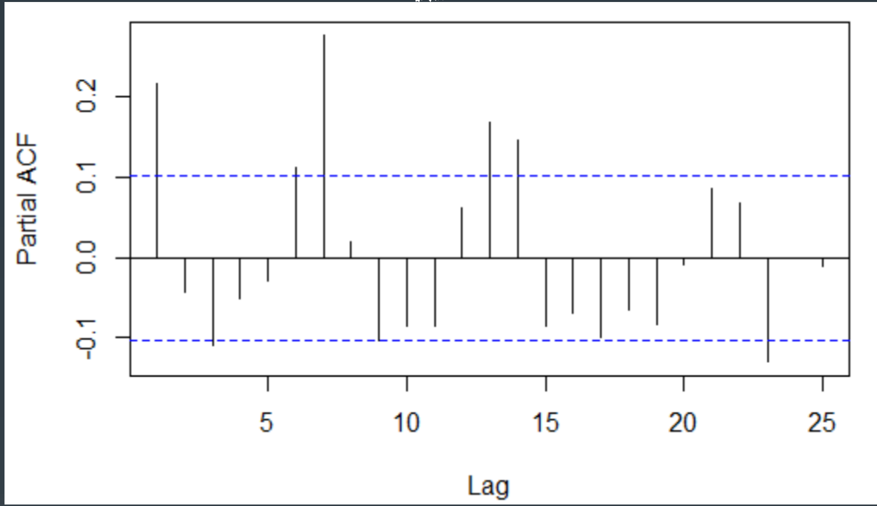




- Chrome is also popular on Mac because it contributes more revenue than Windows from operating system category

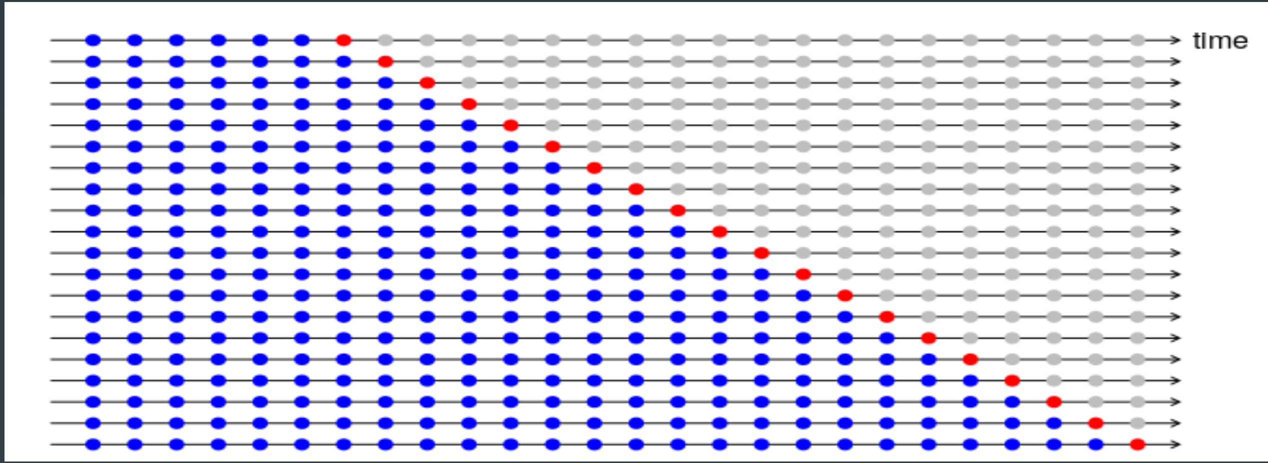
Time Series Model





- Based on the results of ACF and PACF plots, as well as the discovery of EDA part, the Frequency of the data should be 7.
- The time series modeling requires a transformation on data to make it more stationary

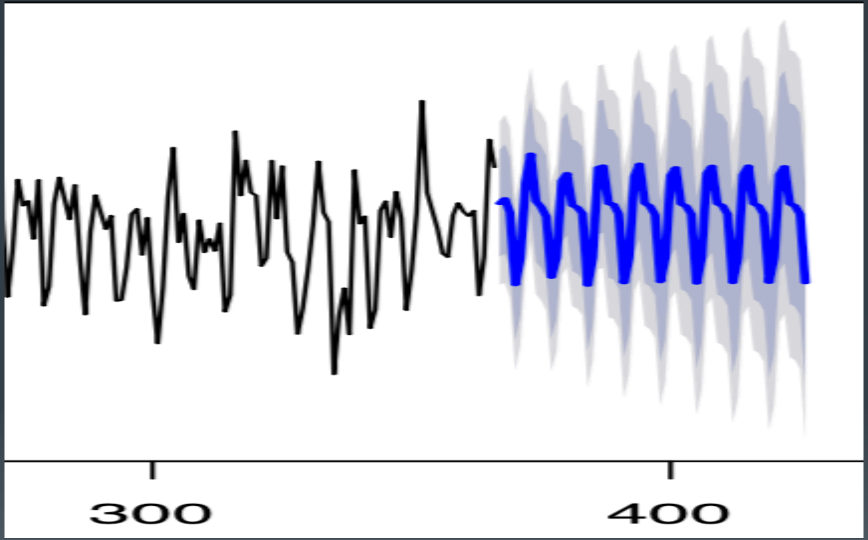
Time Series Cross Validation



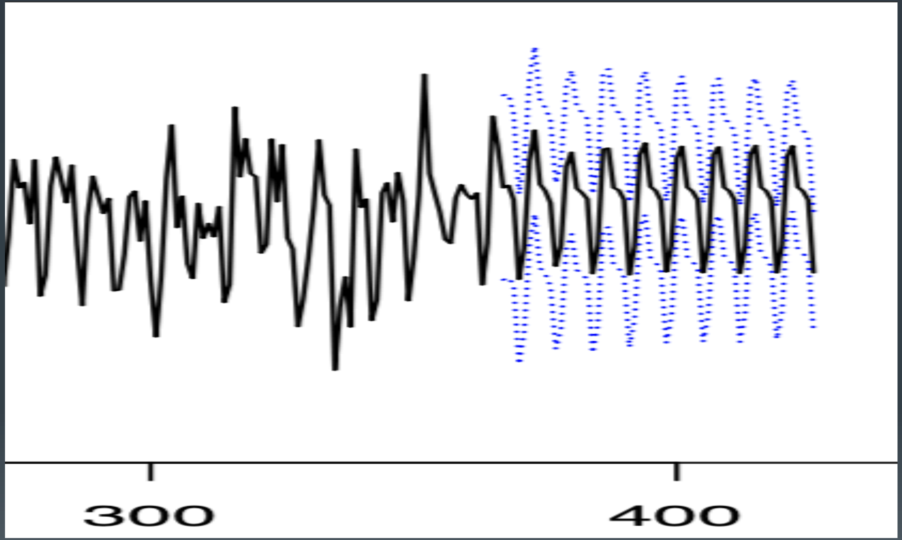
- We fitted a SARIMA model according to the best AIC
- Step by step RMSE 2.291629; Whole set RMSE 2.227386

Time Series Forecasting

- Basic Fixed Window Result



- Rolling Window Result



Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	…
5052.00	5200.40	4303.18	1171.33	2078.32	7036.61	9876.81	…

Classification Models

- Logistic

Act / Pred	0	1
0	219,536	1,081
1	3,499	1,798

- Random Forest

Act / Pred	0	1
0	218,201	128
1	4,834	2,751

- XGBoost

Act / Pred	0	1
0	218,386	141
1	4,649	2,738

- The ratio of Transaction Revenue made and No Revenue made has been rebalanced into 1:10 instead of 1:50 in train dataset
- All models have been tested in original test dataset

Classification Models

	Accuracy	Precision	Recall	F-1 Score
Logic	97.97%	33.94%	62.45%	43.98%
Random Forest	97.80%	36.26%	95.55%	52.58%
XGBoost	97.88%	37.07%	95.10%	53.34%

- Train dataset has been rebalanced into 1:10
- Random Forest and XGBoost have very close values; both can be used for prediction.

Linear Mixed Models

- LMM 0: Revenue ~ (1| Visitor Id)
- LMM 2: Revenue ~ page views + (1| Visitor Id)
- LMM 3: Revenue ~ page views + hits + (1| Visitor Id)
- LMM 4: Revenue ~ page views + hits + visit Number + (1| Visitor Id)

……..

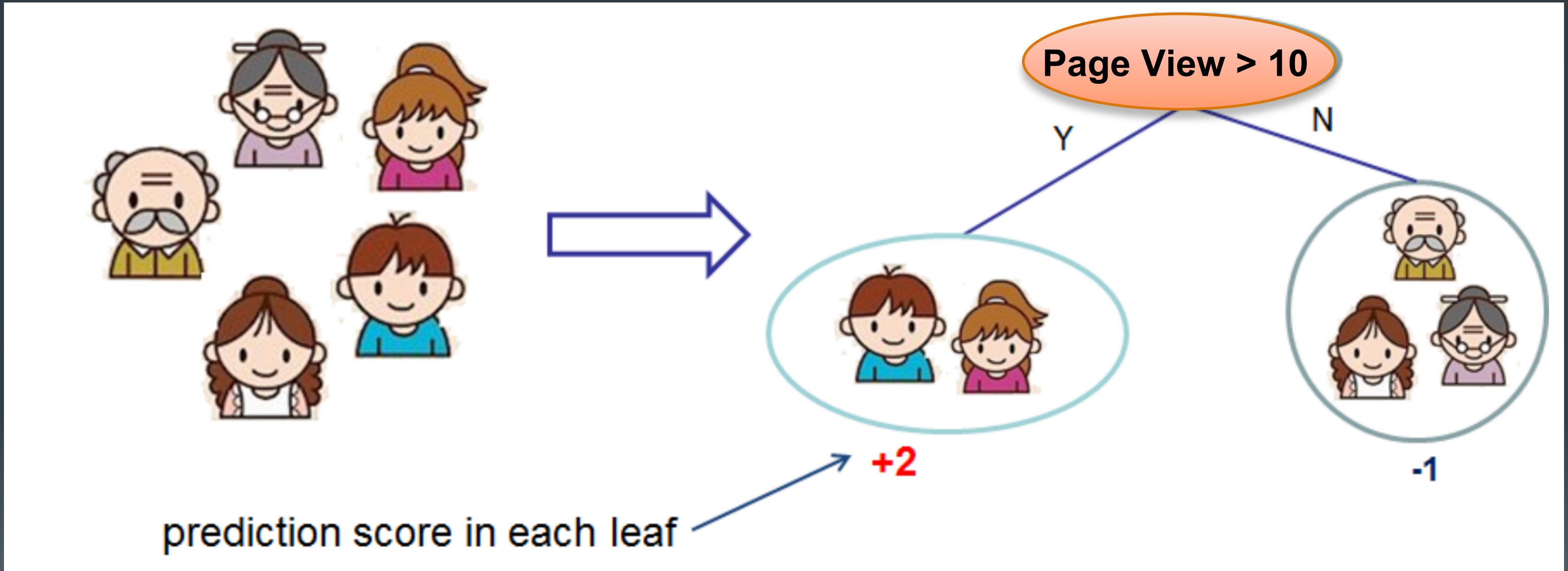
- LMM 7: Revenue ~ page views + hits + visit Number + channel Grouping + browser + operating System + country + (1| Visitor Id))

Linear Mixed Models

	DF	AIC	BIC	LOGLik	deviance	Chisq	P-value
LMM 0	3	26,940	26,961	-13,467	26,934		
LMM 1	4	26,495	26,524	-13,244	26,487	446.5118	$< 2.2e^{-16}$
LMM 2	5	26,332	26,368	-13,161	26,322	165.1524	$< 2.2e^{-16}$
LMM 3	6	26,287	26,330	-13,138	26,275	46.8583	$7.631e^{-12}$
LMM 4	7	26,287	26,337	-13,137	26,273	2.0868	0.1486
LMM 5	8	26,133	26,190	-13,059	26,117	156.1888	$< 2.2e^{-16}$
LMM 6	9	**26,130**	26,193	-13,056	26,112	5.5482	0.0185
LMM 7	10	26,131	26,202	-13,056	26,111	0.5196	0.4710

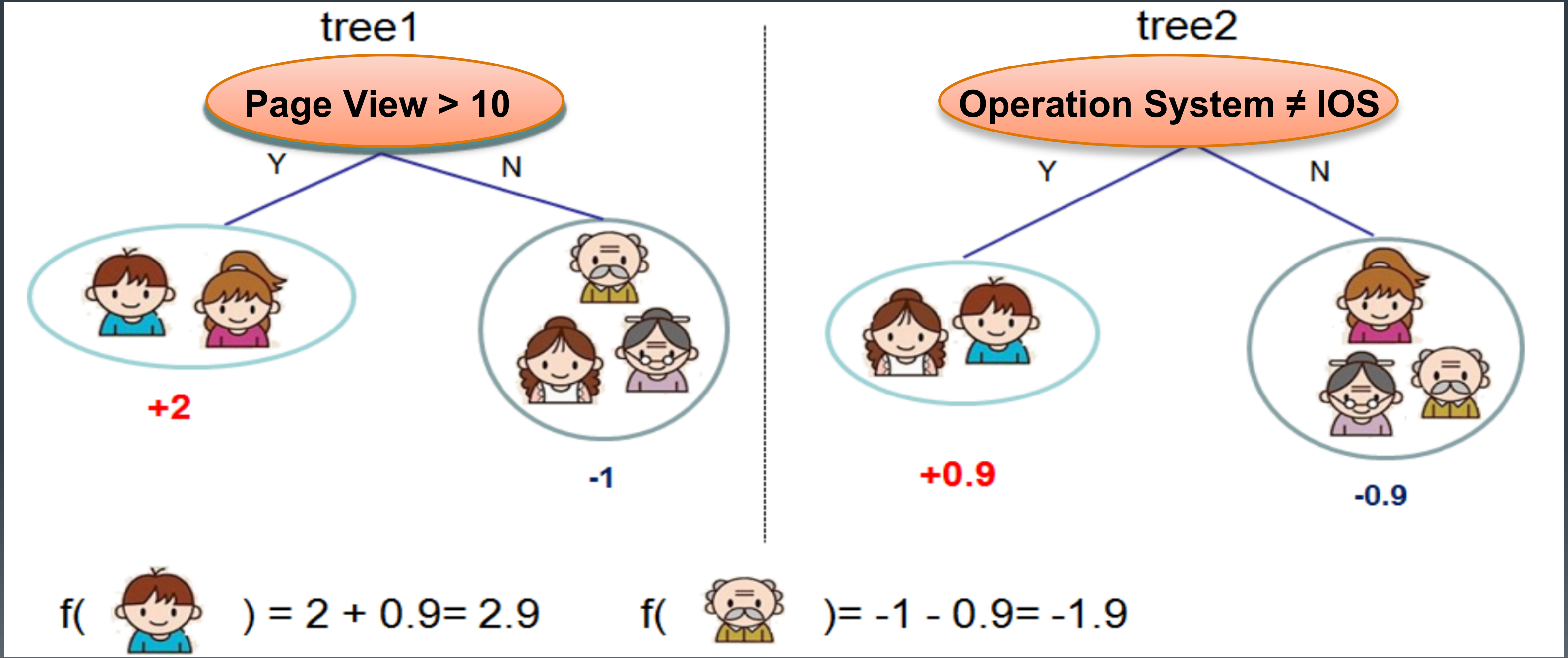
- Select the best model based on the lowest AIC

XGBoost



- This graph is **SOOOO CUTE** !!

XGBoost



- This graph is **SOOOO CUTE** !!

RMSE Table

- Evaluation Method: $$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2},$$

- Evaluation Table:

	Train	Test
LMM	0.5878892	1.05369
XGBoost	1.016699	1.065387

Overfitting

- Further thinking: the LASSO or Ridge regression

Conclusion

- SARIMA $(1,0,1) * (2,1,0)_7$ would produce the best prediction on the daily average revenue in the future 60 days
- Random Forest and XGBoost would more accurately determine customer behavior
- XGBoost would be evaluated as the best model by RMSE on predicting transaction revenue amount

If we have more time …

- Improvement on data balancing
- Too many levels of categorical data: Potentially drop column
- Improvement on dimension reduction
- Confusion matrix trade-off

Questions?

Refence

1. Google Merchandise Store, shop.googlemerchandisestore.com/.

2. Google Analytics Customer Revenue Prediction, Google, Dec. 2017, www.kaggle.com/c/ga-customer-revenue-prediction/overview/description.

3. Hyndman, Rob J, and George Athanasopoulos. Forecasting: Principles and Practice. Monash University, 2016.Chapter: Evaluating forecast accuracy

4. “Introduction to Boosted Trees.” XGBoost, 2016, xgboost.readthedocs.io/en/latest/tutorials/model.html.