Using R for Regulatory Stress Testing Modeling

Thomas Zakrzewski (Tom Z.,)
Head of Architecture and Digital Design
S&P Global Market Intelligence
Risk Services

May 19th, 2017

S&P Global
Market Intelligence

Project Overview

- 2008 Financial meltdown due to asset securitization and overleveraging
- Response in form of regulatory requirements for Bank Holding Companies
  - Dodd-Frank Act supervisory stress testing (DFAST) – BHCs of $10-$50 BN Total Assets must provide forward-looking stress tests of their capital structure in-house.
  - Comprehensive Capital Analysis and Review (CCAR) – Further to DFAST requirements, BHCs of more than $50 BN Total Assets are also subject to Fed-conducted stress tests which must be publicly-disclosed.
- Both programs assess whether:
  - BHCs possess adequate capital to sustain macro and market shocks while still meeting lending needs without need of government capital injections
  - Capital positions fall below ratio thresholds under 3 hypothetical scenarios: Baseline, Adverse, and Severely Adverse
- Using R for Data Exploratory and Modeling

S&P Global Market Intelligence – Risk Services

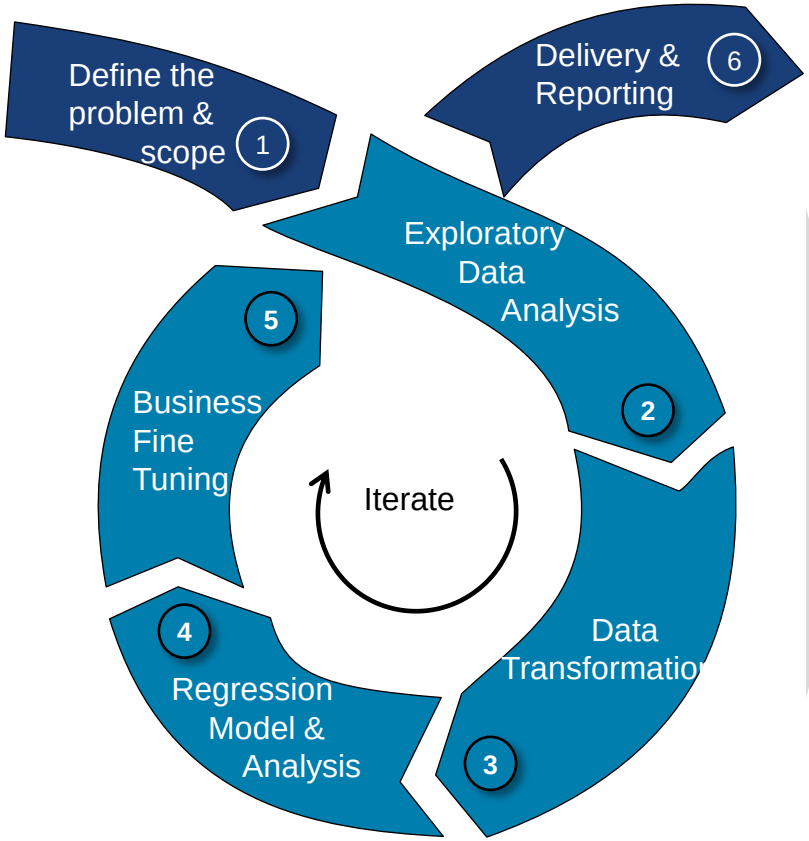
- S&P Global Market Intelligence combines broad data, powerful analytics, and deep sector intelligence to give our clients unrivaled insight into the companies and markets they follow.

- Risk Services Provide essential data, tools, and analytical models for credit and risk management professionals needed to identify and manage potential default risks of private, publicly traded, rated and unrated companies (obligors) of any size, across a multitude of sectors globally

- Working with Educational Institutions
  - Capstone Project with Columbia Business School – Regulatory Stress Test Models
    - Market Intelligence - Eduardo Alves, Yuri Katz and Thomas Zakrzewski
    - Columbia Business School - Students: Maxime Bourgeon, Yu-Cheng Chang, Yufei Chen, Gabriel Lerner, Yueran Li, Víðir Þór Rúnarsson, Sonalika Sangwan and Jinxian Yang under consultation of Professor Souleymane Kachani

Methodology

The project followed multiple iterations of exploratory data analysis, data transformation and imputation, modeling & analysis, testing and business fine tuning



1 Define the Problem & Scope

- Based on regulatory Call Report data estimate expected losses in each loan category based on: Default Rate Model, Lost Rate Model, Portfolio Growth Model

2 Exploratory Data Analysis

- As a first step to understanding data, correlation matrix of predictors was calculated and Q-Q plots to test assumption that the data is normally distributed

3 Data Transformation

- Transform data using techniques such as logarithm, standardization and lagged transformation
- Including macroeconomic data (predictors) and PD, LGD, growth rate and PPNR (responses) at portfolio level

4 Regression Model and Analysis

- Regress the response against the predictors (regressors) and select the variables using techniques such cutoff at predefined p-value, stepwise, forward, Lasso regression & ARIMA
- Choose the model based on business knowledge and other statistical criterion such as AIC, BIC and adjusted R-square

5 Business Fine Tuning

- Check point to ensure the model make business sense
  - Are the variables selected by the model associated with the portfolio that is being projected?
  - Are the projected values sensible compared to historical data and macroeconomic data?

Data Description

The data was transformed into panel format by aggregating data points for all banks as well as the macro-economic data

Bank Financial Statement Data from MI Platform

- Institutions: 31 banks subject to CCAR
- Time frame: quarterly financials from 2001 Q1 to 2016 Q3
- Total number of observations: 1984
- Financial data including:
  - Loan Balances and Charge-offs
  - Securities (AFS and HTM)
  - Equity
  - Capital
  - Pre-Provision Net Revenue

Macroeconomic Data from Federal Reserve

Historical data:

- 16 historical macroeconomic variables
- Time frame: 2001 Q1 to 2016 Q3
- Number of observations: 63

Scenario data:

- 3 scenarios: baseline, adverse, and severely adverse
- Time frame: 2016 Q4 to 2019 Q1
- Number of observations: 10

Final Panel Data

	Default Rate (First Lien Mortgage)	Loss Rate (First Lien Mortgage)	Growth Rate(First Lien Mortgage)	…	Nominal GDP	Nominal Disposable Income	…
2001 Q1				…			…
2001 Q2				…			…
…	…	…	…	…	…	…	…

S&P Global
Market Intelligence

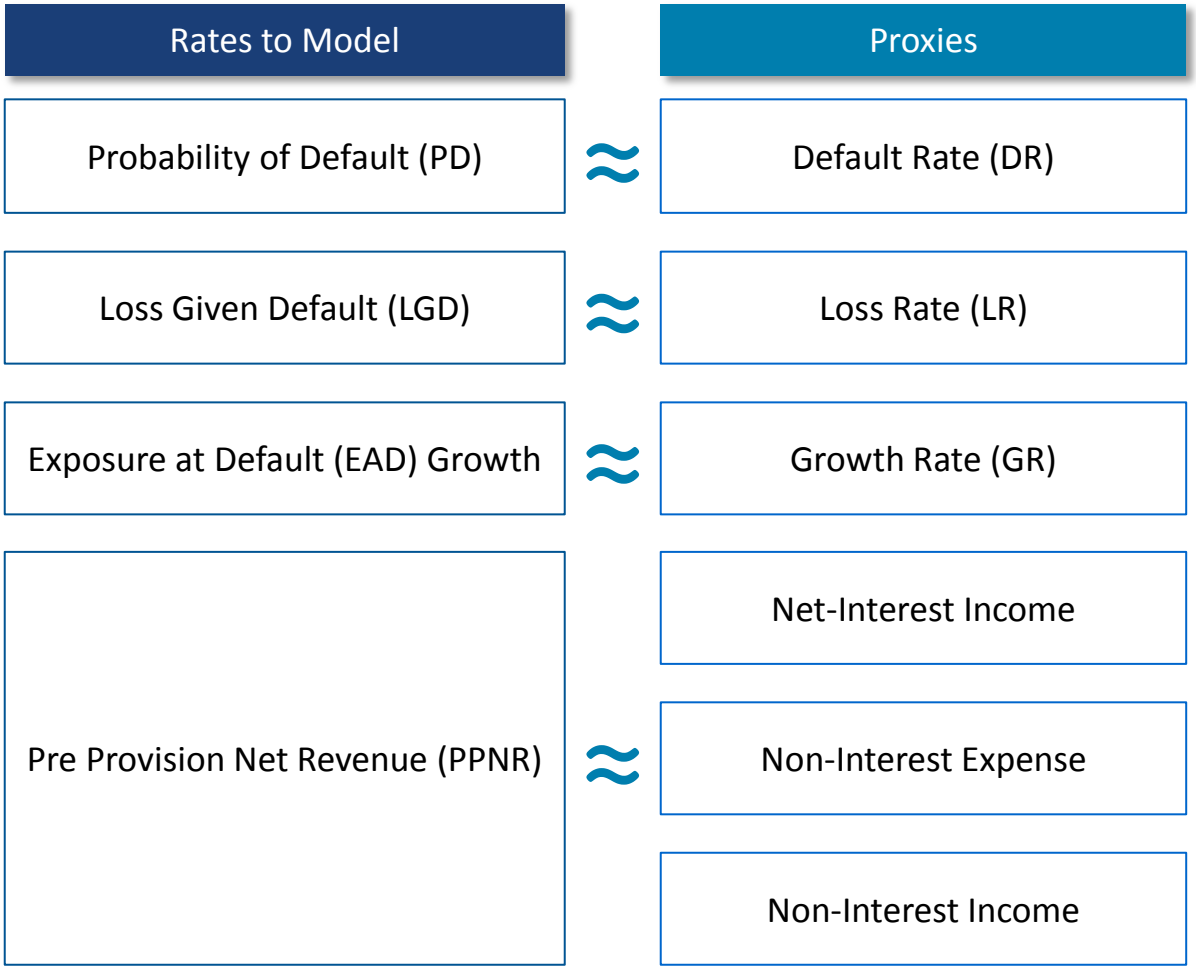
Data Description

There are 15 loan types in the credit risk component of call report

	Loan Type
1	First lien mortgages
2	Closed-end junior liens
3	HELOC (home equity line of credit)
4	C&I loans (commercial & industrial)
5	1-4 family construction loans
6	Other construction loans
7	Multifamily loans
8	Non-farm, non-residential owner occupied loans
9	Non-farm, non-residential other loans
10	Credit cards
11	Automobile loans
12	Other consumer
13	All other loans & leases
14	Loans covered by FDIC loss sharing agreements
15	Total loans & leases

Data Description – Dependent Variables

We used the following proxies to model the corresponding rates



S&P Global
Market Intelligence

Variable Selection & Regression Models

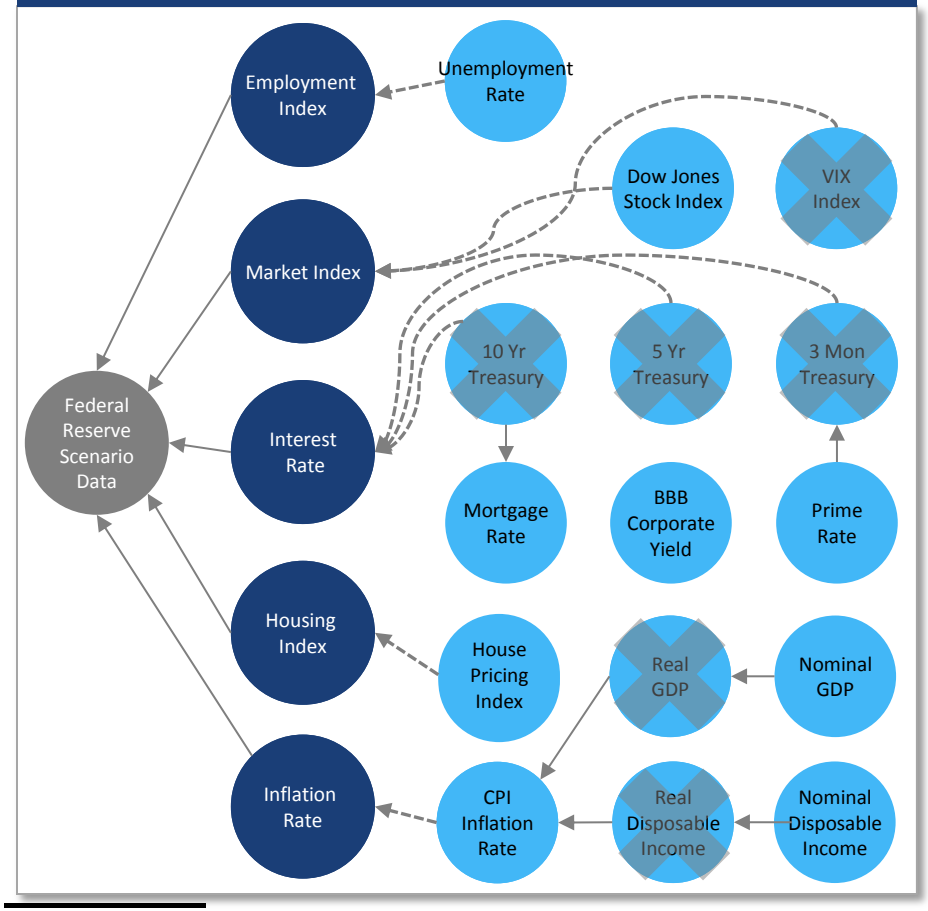
ARIMA has been chosen to be the champion model: remaining models violated iid principle (independent and identically distributed) while error terms showed strong autocorrelations

Stepwise	Lasso Regression	ARIMA Model ✔
Key Assumptions • The data follows linear relationship between the responses and predictors • Residuals are normally distributed	**Key Assumptions** • The data sample follows linear relationship • There is no outliner that will influence estimation of parameters • Residuals are normally distributed	**Key Assumptions** • The predictors follow normal distribution and could be normalized • There exists time series patterns with auto-correlated terms • The periodicity of PD & LGD is 4 quarters
Advantages • Regression model is robust and can fit the data even if some assumptions are violated • Simple approach suitable for first exploratory iteration • Selection results are easier to interpret	**Advantages** • Model shrinks the coefficients of variables so the result is more business interpretable • Selection results are easier to interpret	**Advantages** • Capable of capturing the general trend and the time-series trend • High-level method analyzing the historical data better, yielding better results in general • Results are easier to interpret
Weaknesses • Incapable of capturing time-series characteristics • The model keeps excess number of variables (overfitting), also leaving it hard for business side to find proper business logic to explain the result	**Weaknesses** • Incapable of capturing time-series characteristics • Over-shrinkage: the model wipes out all coefficients of independent variables for some portfolio which results in ill-prediction	**Weaknesses** • Certain portfolios require manual variable selection with business expertise • ARIMA's precision is affected by the data completeness (it could perform better with longer time span)

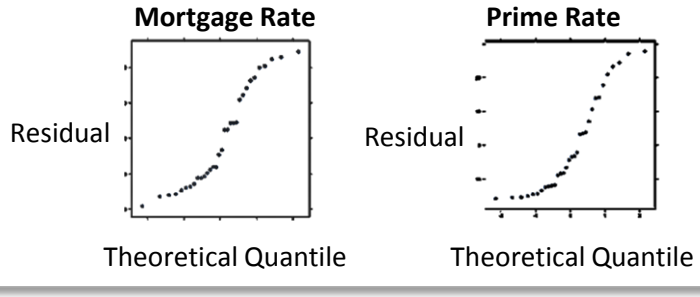
Exploratory Data Analysis and Data Transformation

7 macroeconomic variables were firstly removed from model due to high correlation with other variables; normalization was performed on all remaining 9 variables

Macroeconomic Variable Selection (Predictors)



Selected Quantile-Quantile Plots



Transformations & Key Assumptions

Assumptions:

- All remaining variables are relatively normally distributed according to the QQ residual plots
- Mortgage rate, BBB corporate yields and Prime rate can represent 3 types of treasury yields in regression

Selection & Transformation:

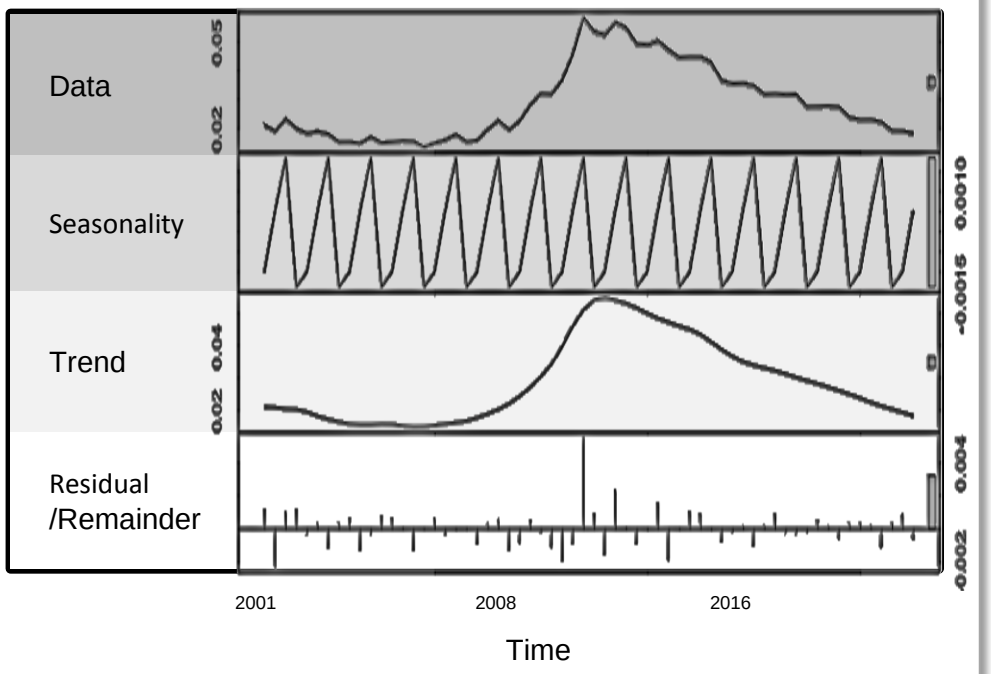
- Normalize all remaining variables across years from 2001 to 2020
- Add a lagged term (a quarter) of unemployment rate to the list of variables

S&P Global
Market Intelligence

Linear Regression with ARIMA Model

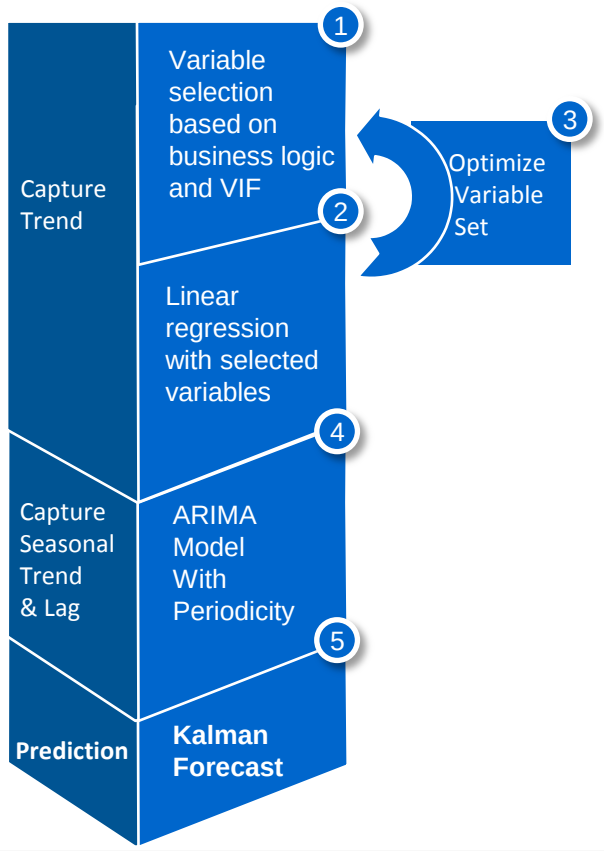
In the Linear Regression with ARIMA Model, fit using linear regression firstly to capture trend; then ARIMA on residuals; finally, forecast using Kalman Filter

Decomposition of LGD Total Loan & Leases Data



Decomposition of data shows us the trend, seasonality and reminder of LGD total loan and leases data

Schematic of ARIMA Model



S&P Global
Market Intelligence

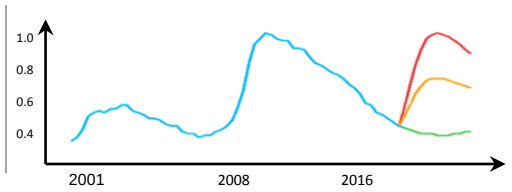
Sample Model Result – Multifamily Loans (DR)

Unemployment rate and BBB Corporate Yield are chosen to be the independent variables in modeling Probability of Default

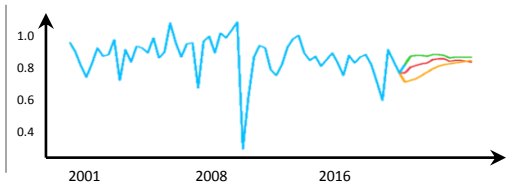
Adverse Scenario
Severely Adverse Scenario
Historical Data Points
Baseline Scenario

Independent Variables

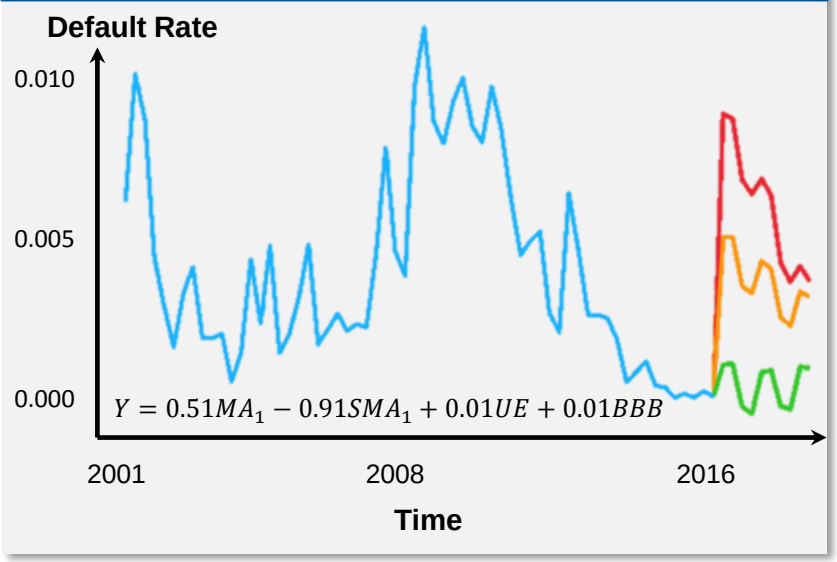
Unemployment Rate



BBB Corporate Yield



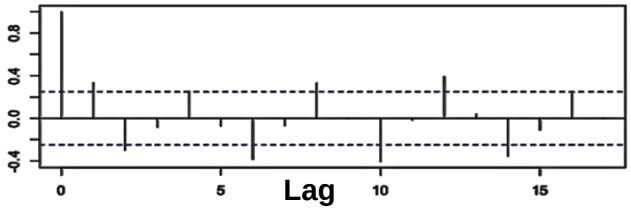
Default Rate (LR) Prediction



Ljung-Box Test Result

- p-value for Ljung-box Test result
  - 0.00731 < 0.05
- This indicates the possibility of non-zero autocorrelation

Auto Correlation Function of ARIMA Residuals



S&P Global
Market Intelligence

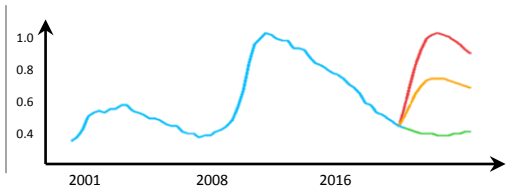
Sample Model Result – Multifamily Loans (LR)

Unemployment rate (Lag 1) and Dow Jones Stock Index are chosen to be the independent variables in modeling Loss Given Default

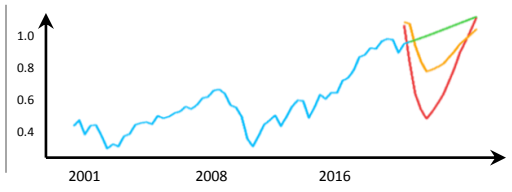
Adverse Scenario
Severely Adverse Scenario
Historical Data Points
Baseline Scenario

Independent Variables

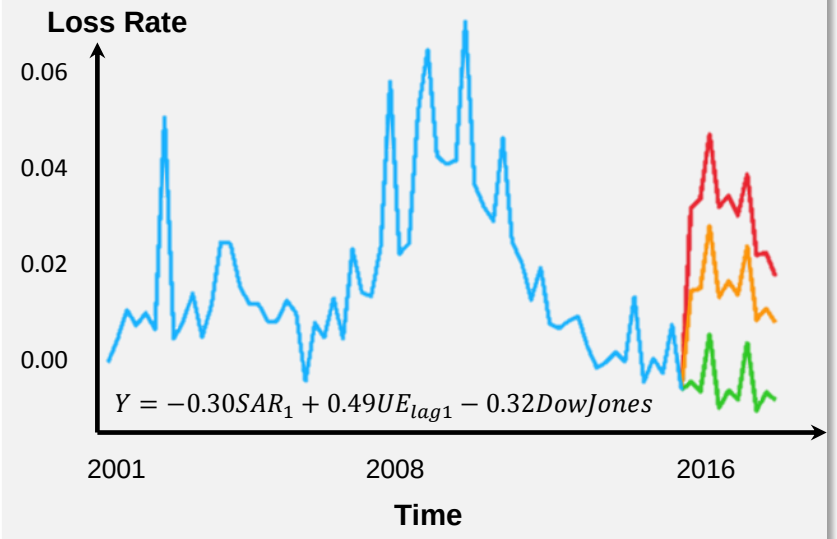
Unemployment Rate (Lag 1)



Dow Jones Stock Index



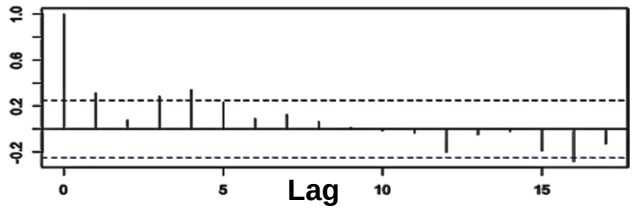
Loss Rate (LR) Prediction



Ljung-Box Test Result

- p-value for Ljung-box Test result
  - 0.01298 < 0.05
- This indicates the possibility of non-zero autocorrelation

Auto Correlation Function of ARIMA Residuals

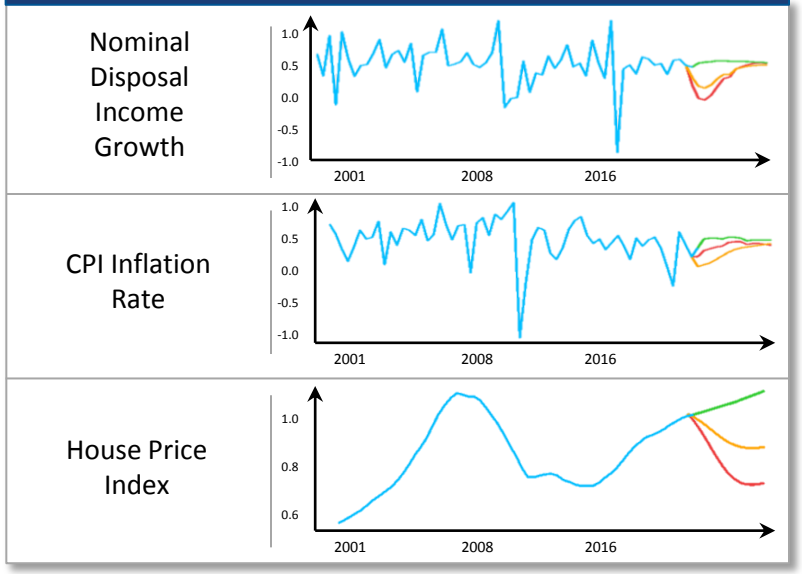


S&P Global
Market Intelligence

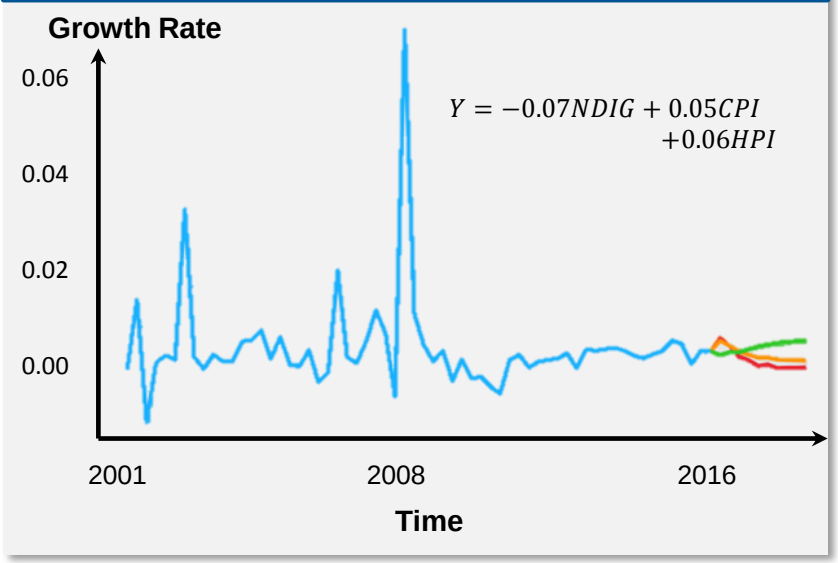
Sample Model Result – Multifamily Loans (GR)

Nominal Disposal Income Growth, CPI Inflation Rate, and House Price Index are chosen to be the independent variables in modeling Growth Rate

Independent Variables



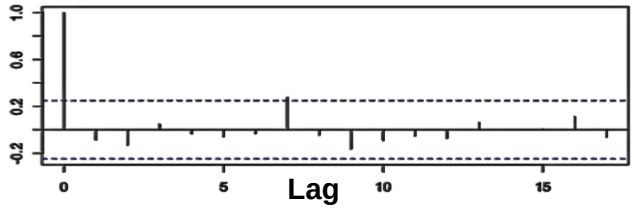
Growth Rate (LR) Prediction



Ljung-Box Test Result

- p-value for Ljung-box Test result
  - 0.47801 > 0.05
- This indicates no auto-correlation in the residual

Auto Correlation Function of ARIMA Residuals



S&P Global
Market Intelligence

Conclusion

Overall, the project results are promising and it is recommended to further develop the prototype; going forward, data incompleteness should be taken into consideration

Challenges

- Working with small data set
- Lack of complete historical data and small number of data points
- Making key assumptions
- Choosing proxies for modeling
- Enforce seasonal structure on PD & LGD Model
- Evolving Regulatory Landscape
- New efforts to deregulate banks could change modeling requirements and needs

Learning for Columbia Team

- Data science topics
  - Data Transformation & Imputation
  - Variable Selection Framework
  - Exploratory Data Analysis
  - Time Series Analysis in R
- Credit risk management topics
  - Stress Testing general knowledge
  - Corporate credit risk analysis

Thank you

Thomas Zakrzewski (Tom Z.,)

Head of Architecture and Digital Design

S&P Global Market Intelligence

Risk Services

T: 212.438.8458

Thomas.Zakrzewski@spglobal.com