Data Science Final Project

VICTOR TANG

05 AUG 2017

What is the project about?

An Investigation on the association between financial asset price and macro-economic factors, company-specific factors and other asset prices.

Financial asset of interest in this project
– Hong Kong-listed equity share of Industrial and Commercial Bank of China (ICBC)

Project Summary (1)

Problem Statement - How is the Hong Kong-listed share price of ICBC associated with the price of a number of other assets as well as a few macro-economic and company-specific factors?

Null Hypothesis - The share price of ICBC has no association with the price of the few assets or any of the macro-economic or company-specific factors under investigation.

Alternative Hypothesis - The share price of ICBC is associated with the price of a number of other assets and/or a few macro-economic and company-specific factors.

Project Summary (2)

Dataset:
- For the dependent variable, I have acquired 5 years of data of ICBC share price from Bloomberg (Jul 2012 - Jul 2017).
- For the independent variables, I have also acquired the same 5 years of data on interest rates, interest rate differentials, equity index levels, foreign exchange spot levels, M2 growth YoY, New RMB Loans, company quarterly earnings, Tier 1 capital ratios, Non-Performing Loans (NPL) growth YoY from Bloomberg.

Data Dictionary	
Variable Name	Meaning
ICBC_Log_Ret (Dependent Variable)	Daily log return of ICBC share's price in decimals, 0.01 is 1% daily log return
IR_Chg_1Y	Daily change in 1-year interest rate (ir) in decimals, 0.01 is 1% change in ir
IR_Chg_5Y1Y	Daily change in (5-year ir - 1-year ir) in decimals, 0.01 is 1% change in ir
HSCEI_Log_Ret	Daily log return of HSCEI Index's level in decimals, 0.01 is 1% daily log return
USDCNH_Log_Ret	Daily log return of USDCNH spot price in decimals, 0.01 is 1% daily log return
M2_Growth_YoY	China's year-on-year (YoY) M2 growth rate in decimals, 0.01 is 1% growth rate
New_Loans_Growth_YoY	China's YoY new loans growth rate in decimals, 0.01 is 1% growth rate
Earning_Growth_YoY	ICBC's YoY earning growth rate in decimals, 0.01 is 1% growth rate
T1_Cap_Ratio_Chg	ICBC's YoY change in tier-1 capital ratio in decimals, 0.01 is 1% change
NPL_Growth_YoY	ICBC's YoY non-performing loan growth rate in decimals, 0.01 is 1% growth rate

Project Summary (3)

Method and Model:
I attempted to use a multivariate linear regression model, with the daily return of the share price of ICBC as the dependent variable and the daily return of a few other assets and the periodical change of a few macro-economic and company-specific factors as the independent variables.

I divided the data into a training set and a test set with an 80%/20% breakdown. I used the 20% of data in the test set to gauge the efficacy of the predictions made based on the model and parameters generated from the 80% of data in the training set.

Bloomberg Snapshots

Data after Clean-up and Date Matching

|	A	B	C	D	E	F	G	H	I	J	K
1	Date	ICBC_Log_Ret	IR_Chg_1Y	IR_Chg_5Y1Y	HSCEI_Log_Ret	USDCNH_Log_Ret	M2_Growth_YoY	New_Loans_Growth_YoY	Earning_Growth_YoY	T1_Cap_Ratio_Chg	NPL_Growth_YoY
2	7/20/2012	0.0167	-0.0002	0	0.0058	0.0005	0.1364	0.451	0.152	0.0039	0.0317
3	7/23/2012	-0.0216	0.0004	-0.0007	-0.0317	0.0019	0.1364	0.451	0.152	0.0039	0.0317
4	7/24/2012	-0.0024	-0.0001	0.0006	-0.0059	0.0002	0.1364	0.451	0.152	0.0039	0.0317
5	7/25/2012	0.0073	0	-0.0004	0.0002	0.0003	0.1364	0.451	0.152	0.0039	0.0317
6	7/26/2012	-0.0048	0.0002	-0.0003	-0.0009	-0.0004	0.1364	0.451	0.152	0.0039	0.0317
7	7/27/2012	0.0263	0.0003	0.0005	0.0202	-0.0007	0.1364	0.451	0.152	0.0039	0.0317
8	7/30/2012	0.0187	0.0003	-0.0003	0.0132	0.0001	0.1364	0.451	0.152	0.0039	0.0317
9	7/31/2012	0.0296	0	0.0002	0.0157	-0.0011	0.1394	0.0987	0.152	0.0039	0.0317
10	8/1/2012	0.0156	-0.0002	0.0001	0.0091	0.0004	0.1394	0.0987	0.152	0.0039	0.0317
11	8/2/2012	-0.0089	0	-0.0004	-0.0096	-0.0002	0.1394	0.0987	0.152	0.0039	0.0317
12	8/3/2012	0.0045	0	0.0002	-0.0009	0.0008	0.1394	0.0987	0.152	0.0039	0.0317
13	8/6/2012	0.0132	0.0013	-0.0001	0.0156	-0.0005	0.1394	0.0987	0.152	0.0039	0.0317
14	8/7/2012	0.0044	0.0003	0	0.0039	-0.0007	0.1394	0.0987	0.152	0.0039	0.0317
15	8/8/2012	-0.0044	0	0	0.0016	-0.0005	0.1394	0.0987	0.152	0.0039	0.0317
16	8/9/2012	0.0044	-0.0002	-0.0002	0.0095	-0.0008	0.1394	0.0987	0.152	0.0039	0.0317
17	8/10/2012	-0.0154	-0.0005	-0.0001	-0.0057	0.0004	0.1394	0.0987	0.152	0.0039	0.0317
18	8/13/2012	-0.0044	0.0008	-0.0001	-0.0092	-0.0004	0.1394	0.0987	0.152	0.0039	0.0317
19	8/14/2012	0.0133	0.0005	-0.0001	0.0102	-0.0003	0.1394	0.0987	0.152	0.0039	0.0317
20	8/15/2012	-0.011	0.0006	-0.0004	-0.0138	0.0011	0.1394	0.0987	0.152	0.0039	0.0317
21	8/16/2012	-0.0134	0	0.0002	-0.0038	-0.0001	0.1394	0.0987	0.152	0.0039	0.0317
22	8/17/2012	0.0112	-0.0002	-0.0001	0.0091	-0.0006	0.1394	0.0987	0.152	0.0039	0.0317
23	8/20/2012	-0.0089	0.0017	-0.0004	-0.0037	-0.0007	0.1394	0.0987	0.152	0.0039	0.0317
24	8/21/2012	0.0045	-0.0001	0.0001	0.0032	-0.0008	0.1394	0.0987	0.152	0.0039	0.0317
25	8/22/2012	-0.0135	0.0015	-0.0012	-0.013	0.0002	0.1394	0.0987	0.152	0.0039	0.0317
26	8/23/2012	0.0157	-0.0009	0.0002	0.0141	0	0.1394	0.0987	0.152	0.0039	0.0317
27	8/24/2012	-0.0203	-0.0002	0.0002	-0.0165	0.0002	0.1394	0.0987	0.152	0.0039	0.0317
28	8/27/2012	-0.0184	0	-0.0002	-0.0135	0.0005	0.1394	0.0987	0.152	0.0039	0.0317
29	8/28/2012	0.0046	0.0002	-0.0006	-0.0024	-0.0003	0.1394	0.0987	0.152	0.0039	0.0317
30	8/29/2012	-0.014	0	0.0004	-0.0054	-0.0005	0.1394	0.0987	0.152	0.0039	0.0317
31	8/30/2012	-0.0189	0.0005	-0.0004	-0.0138	-0.0002	0.1394	0.0987	0.152	0.0039	0.0317
32	8/31/2012	0.0024	0	-0.0001	-0.0065	-0.0006	0.1346	0.2836	0.152	0.0039	0.0317
33	9/3/2012	0	0.0008	-0.0002	0.0008	-0.0006	0.1346	0.2836	0.152	0.0039	0.0317

Exploratory Data Analysis (EDA) / Modeling Preparation
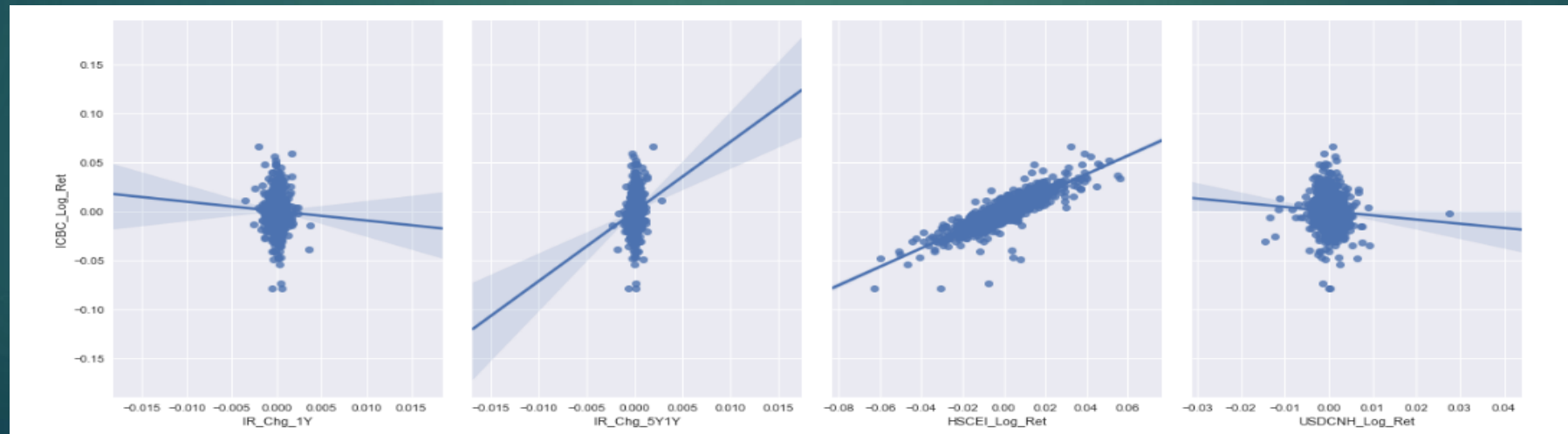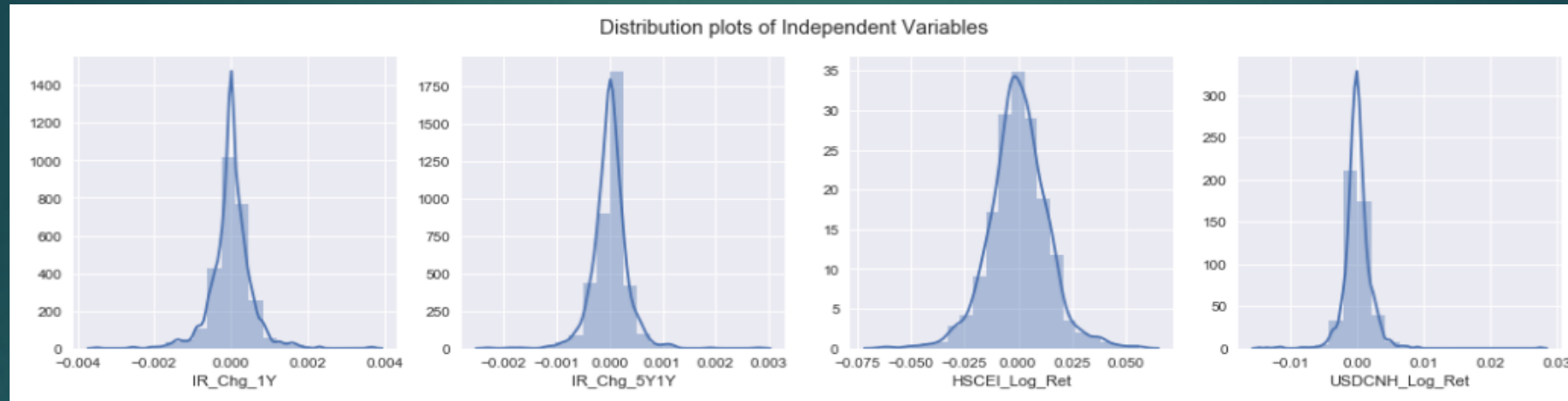
	ICBC_Log_Ret	IR_Chg_1Y	IR_Chg_5Y1Y	HSCEI_Log_Ret	USDCNH_Log_Ret	M2_Growth_YoY
ICBC_Log_Ret	1.000000	-0.034661	0.156918	0.870885	-0.060591	-0.022820
IR_Chg_1Y	-0.034661	1.000000	-0.357682	-0.021264	-0.016022	0.051528
IR_Chg_5Y1Y	0.156918	-0.357682	1.000000	0.177006	-0.016581	-0.033022
HSCEI_Log_Ret	0.870885	-0.021264	0.177006	1.000000	-0.059301	-0.025791
USDCNH_Log_Ret	-0.060591	-0.016022	-0.016581	-0.059301	1.000000	-0.022410
M2_Growth_YoY	-0.022820	0.051528	-0.033022	-0.025791	-0.022410	1.000000
New_Loans_Growth_YoY	-0.065739	0.008503	-0.004501	-0.043422	0.049915	-0.009178
Earning_Growth_YoY	0.043155	0.029066	-0.003012	0.036459	-0.047603	0.501407
T1_Cap_Ratio_Chg	-0.013082	-0.021515	-0.017993	-0.023278	0.012794	-0.053860
NPL_Growth_YoY	-0.060628	-0.044362	-0.009313	-0.060241	0.050776	-0.074634
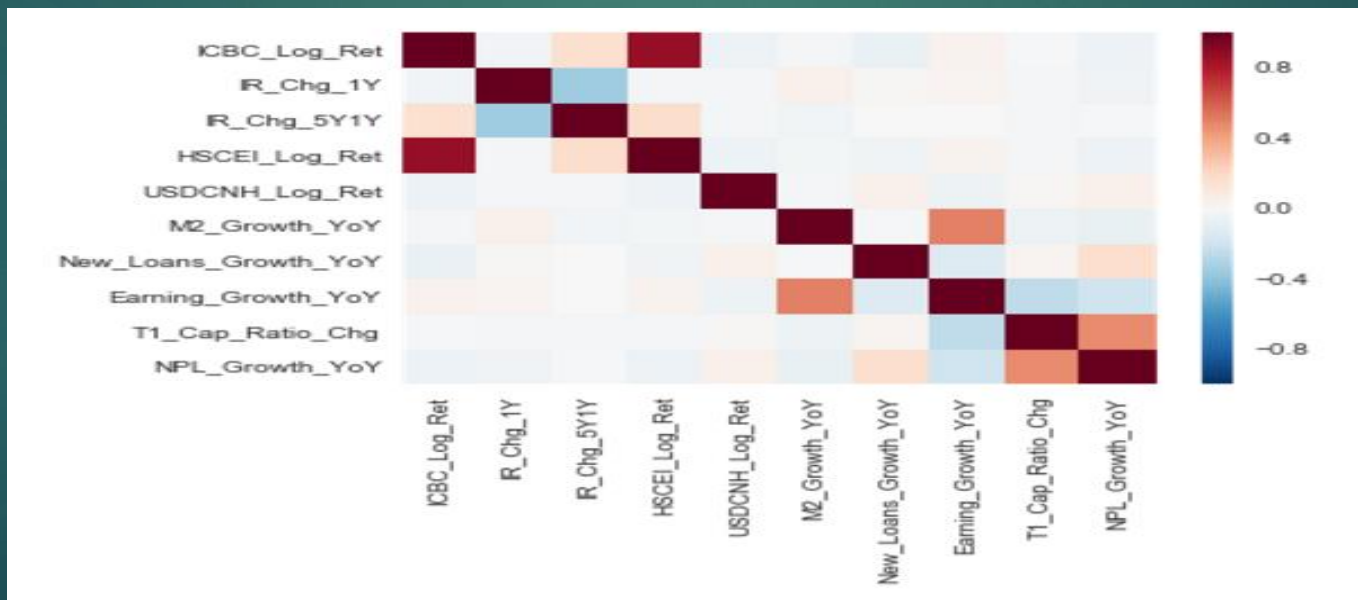
Findings from EDA:

The dependent variable and most of the independent variables have distributions that are close to a normal distribution (although some have a certain degree of skew). Only the quarterly data exhibit less of such a feature which could be due to an insufficient number of data points.

The visualization shows that the dependent variable is more correlated to a few independent variables than others, the two independent variables that stand out are HSCEI_Log_Ret and IR_Chg_5Y1Y.

From the scattor plot, it further shows that the dependent variable has a relatively high degree of linear association with the independent variable HSCEI_Log_Ret.

Modeling – Multivariate Linear Regression (Ordinary Least Squares)

mlr = smf.ols(formula= "ICBC_Log_Ret ~ IR_Chg_1Y + IR_Chg_5Y1Y + HSCEI_Log_Ret + USDCNH_Log_Ret + M2_Growth_YoY + ..., data=data_train).fit()

Regression Output:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0029	0.002	1.330	0.184	-0.001	0.007
IR_Chg_1Y	-0.1565	0.469	-0.334	0.739	-1.076	0.763
IR_Chg_5Y1Y	0.0684	0.759	0.090	0.928	-1.421	1.558
HSCEI_Log_Ret	0.9419	0.017	54.205	0.000	0.908	0.976
USDCNH_Log_Ret	-0.0218	0.111	-0.196	0.844	-0.240	0.196
M2_Growth_YoY	-0.0240	0.017	-1.381	0.168	-0.058	0.010
New_Loans_Growth_YoY	-0.0011	0.001	-1.943	0.052	-0.002	1.09e-05
Earning_Growth_YoY	0.0067	0.004	1.596	0.111	-0.002	0.015
T1_Cap_Ratio_Chg	0.0475	0.051	0.932	0.352	-0.053	0.148
NPL_Growth_YoY	-0.0012	0.005	-0.220	0.826	-0.012	0.009

Model Fitting and Prediction on Training and Test Set

lm = LinearRegression().fit(data_train[factors], data_train["ICBC_Log_Ret"])

…

train_preds = lm.predict(data_train[factors])

…

test_preds = lm.predict(data_test[factors])

Dataset / Prediction Metrics	MAE	MSE	RMSE
Training Set	0.004945	0.0000525	0.007246
Test Set	0.004902	0.0000531	0.007287

K-Fold Validation on the linear regression model (K=10)

RMSE across 10 folds of validation:

RMSE 1	0.006855272
RMSE 2	0.006329371
RMSE 3	0.006430726
RMSE 4	0.009026074
RMSE 5	0.006999241
RMSE 6	0.006500893
RMSE 7	0.009783111
RMSE 8	0.007491569
RMSE 9	0.007043931
RMSE 10	0.005728344
Mean RMSE	0.007218853

Findings from the Analysis (1)

The linear regression analysis shows that 76.1% of the variance in the dependent variable can be explained by the change in the chosen set of independent variables (Adj R-Squared = 0.761).

On the p-value and coefficient CI of the independent variables, only two independent variables (HSCEI_Log_Ret and New_Loans_Growth_YoY) have small p-values and have 95% CI of the coefficient not crossing 0. So these are the two independent variables with the strongest linear association with the dependent variable.

The interpretation of the linear coefficients is that for every 1% increase in the log return of HSCEI index level and 1% increase in the YoY new loans growth rate, the log return of ICBC's share price would increase by 0.9419% - 0.0011% = 0.9408%.

Findings from the Analysis (2)

The Root Mean Squared Error (RMSE) in fitting the data on the training set, the test set and in the k-fold validation process are all very close to each other. This implies that there is no issue of over-fitting the data in our model.

However, the absolute value of the RMSE 0.0072 (0.72%) is relatively large in the context of our prediction on the dependent variable, which could imply that the linear model has not performed very well on prediction on any of the data set.

Further test on a linear model with only two independent variables (HSCEI_Log_Ret and New_Loans_Growth_YoY) shows that RMSE is still around 0.0072 and could not be lowered further due to the limitation on the predictive power of the chosen independent variables.

Conclusion on the Analysis (1)

The overall result of the linear regression analysis seems to have more room for improvement in terms of selecting other variables with more predictive power on the dependent variable.

In this analysis, we have found HSCEI_Log_Ret to be one of the variables that has a high degree of linear association with the dependent variable, other independent variables did not seem to have a significant relationship with the dependent variable (at least during the period under investigation), it looks like we need some other variables like HSCEI_Log_Ret to come up with more meaningful predictions on the dependent variable.

Conclusion on the Analysis (2)

A few potential independent variables that can be added to the linear regression analysis are the log return of some other equity indices, the log return of the share price of certain companies within the same sector as ICBC (i.e. the banking sector), etc.

We can also use leading/lagging data series of the same set of variables to test on the predictive power of the independent variables. The changes in some independent variables such as macro-economic factors might take some time to be reflected in financial asset's price due to the functioning of certain transmission mechanisms.

There are potential higher frequency proprietary data available that can be incorporated into the linear regression model for predictions on the dependent variable.

Further Thoughts and Future Applications:

This project can serve as a start of a research process on finding the association among financial asset prices and a variety of macro-economic and company-specific factors. More comprehensive analysis needs to be done on the selection of variables, adjustment of data series, fitness of models, etc.

After a certain set of variables is proven to be effective in prediction, data on other financial assets in the same sector can be used for further validation of the model. In this way, the model can be turned into a sector-wide prediction model that will work effectively across more financial products.

Thank You!