**Research Proposal**

**Name:** Yuanyuan Kong **Student ID：**5141209033

**Proposed topic of research**

Forecasting Inflation in China – Based on Machine Learning

**Introduction**

Inflation is an important indicator for designing macroeconomic policies. Also, it is closely related to our daily life, that is, we citizens can sense the increase of index prices. Inflation adds uncertainty to our future purchasing power and makes the market inefficient. Both companies and individuals find it difficult to make long-term budget or planning. High inflation can reallocate purchasing power, both nationally and internationally. Many costs can occur with unpredictable inflation, well-accepted costs including cost-pushing, hoarding, social unset and revolts, allocative inefficiencies, shoe leather costs and so on.

Noticing the large impact of inflation on many economic fields, many scholars have tried to predict inflation via different models since the last a few decades. There are numerous techniques available to analyze financial time series data, but those linear models have inherent restrictions. With the development of machine learning, it is possible to use derived algorithms to forecast properties of non-linear models in finance. These models are capable of handling non-linear relationship, random sequences and optimization. However, only few of them are concerned with inflation forecasting and the benchmarks I’ve found are from the USA and India. Therefore, doing researches on this topic via machine learning is of great value.

**Literature review**

Considering lack of research on the same topic, I will try to summarize previous related studies from two perspectives. One is about the critical factors that affect inflation (concerned about choosing the input vector), the other is about previous models to forecast inflation.

**Critical factors influencing inflation**

1. Money supply

Zhang & Zeng (2005) [1] analyze data in China from year 1980 to 2002 and use Cambridge Equation to establish relation between inflation and money supply, and then apply Granger Causality Test to prove that China’s over supply of currency is the cause of increasing CPI.

Zhu (2011) [2] builds econometric models to analyze relationship between inflation and broad money supply based on monthly data in China from October 2008 to November 2010. However, he admits that the broad money supply can only interpret 22.93% of the change in inflation.

1. Exchange rate

Zhang & Yao (2005) [3] use Cointegration approach and error correction model to analyze the relation between the nominal effective exchange rate of Chinese RMB and CPI. They conclude that the exchange rate contributes a lot to the decline of the price level.

Zhang (2009) [4] uses the VAR model, along with the unit Root Test, Cointegration Analysis, Granger Causality Test and Impulse Response Function to analyze the relationship between exchange rate and inflation rate. He finds that in the long term the exchange rate to increase by one for each reference point, the currency 0.003% drop in inflation; Exchange rate with the transfer rate of inflation and the existence of time-delay effect.

1. Real estate price

Wang & He (2005) [5] argue that there exists sound function relationship between the expected rate of return in real estate market and inflation expectation in China.

Wang, Zhu & Tan (2013) [6] uses Laplacian and Bayesian to estimate parameters based on DSGE two-sector model. They show that under the impulse response, the positive impact of real estate demand will lead to a "hump" type of positive response to inflation. Also, fluctuations in the real estate market can amplify the level of inflation and output volatility.

1. Interest rate

Barth & Ramey (2001) [7] use American data to find that the cost channel exists in the US. The increase of the interest rate can give rise to inflation and its impact is remarkable during 1959 to 1979.

Qi & Liu (2011) [8] analyze China’s data from 1997 to 2010, and include interest rate rules in the econometric model for simultaneous calculations shows that the impact of interest rate shocks on inflation through cost channels is significant.

1. Stock price

Gang & Chen (2004) [9] test the Kaul theory and find that the correlation between inflation and stock returns in China is characterized by stages. China's monetary policy shows the inverse economic cycle character between 1995 and 2002, which leads to the negative correlation between stock returns and inflation.

Liu & Wang (2004) [10] illustrate the positive correlation between the volatility of inflation and the level of inflation based on the GRACH model. Then they select the real rate of return of stock index as an independent variable and the adjusted inflation rate values as the dependent variable, and put them into a specific regression model for statistical test. The results show that the volatility of inflation and the stock's real rate of return is negatively correlated.

1. Other factors

Li, Zhang & Luo (2006) [11] base on the Phillips curve which combines the Okun curve to empirically analyze the impact of the oil price changes on the price level in China. They find that the rising oil prices can promote the inflation of our country.

Qu & Jiang (2010) [12] select the annual data of China's economy from 1985 to 2007 and do the empirical test of the aggregate supply function introducing the output gap. Test results show that there is “output gap theory” between China's inflation and economic growth, and China's output gap is positively correlated with inflation.

Wang & Tian (2012) [13] build the SVAR model to illustrate that there is a notable dynamic relationship between broad money supply (M2) growth, inflation and GDP growth.

**Previous models to forecast inflation**

Xiao & Xia (2008) [14] do short-term inflation prediction via the ARIMA model, using the monthly CPI data of China from January 1990 to November 2007. And they find that ARIMA (1,1,10) provides the best forecasts.

Stock & Watson (2013) [15] investigate forecasts of US inflation at the 12-month horizon and discover that the Phillips curve has been more accurate than forecasting models based on other macroeconomic variables, including interest rates, money and commodity prices.

Koop & DK (2012) [16] forecast quarterly US inflation by utilizing the generalized Phillips curve and the econometric method: incorporate dynamic model averaging. The approach allows for temporal coefficients as well as the entire forecasting model to change over time.

Thakur, Bhattacharyya & Mondal (2016) [17] analyze the historical monthly economic data of India between January 2000 and December 2012 and constructed an inflation forecasting model based on feed forward back propagation neural network.

Ülke, Sahin & Subasi (2016) [18] compares different models, including univariate, multi-variate time series models and machine learning models, to predict different indicators of inflation in the US. Empirical tests illustrate that there is no single perfect model for forecasting all the indicators. The SVR does best in forecasting core-PCE inflation. The ARDL provides the lowest prediction error and the highest prediction accuracy for forecasting core-CPI inflation. The machine learning models work better with more volatile and irregular series.

**My thinking**

Machine learning has been widely used to forecast macroeconomic variables, such as exchange rate (nearest neighbor model) [19], volatility of the stock market return (SVR model) [20] Etc. In the field of data mining and data analytics, machine learning is a functional tool especially for generating practical models from complex textual statistics and using the iterated algorithms to optimize the whole loss, thus capable of doing the prediction; in commercial use, this is known as predictive analytics. These analytical models allow researchers to "produce reliable, repeatable decisions and results" and uncover "hidden insights" through learning from historical relationships and trends in the data. [21] Nevertheless, researchers have rarely explored forecast of inflation via machine learning and I’ve found no existing paper in China about the same topic.

I’m going to use results of the last two papers as my baseline. The second paper shows that when it comes to CPI prediction, the autoregressive distributed lag (ARDL) model is better than other models but it can only account for 23% of the CPI. Generative/robust models from machine learning are prone to provide more precise/idealistic outcomes.

There are many models in machine learning and I will list models I want to use in the methodology part. Also, I will not only select the most significant variables as inputs using Chinese data to do the empirical test, but also consider the hysteresis effect of above factors (e.g. Raw inputs contain the data of past five years). I hope I can develop a model most suitable for Chinese inflation forecast to offer useful advices to Chinese government.

**Sources of data**

The inputs are factors which have lagged effects on inflation, and I may add more in the first step of determining significant variables.

I choose the annual inflation rate as the output.

All the above macroeconomic variables can be found in the National Bureau of Statistics of the People’s Republic of China – national data.

Website: <http://data.stats.gov.cn/>

**Methodology**

I don’t know which model can best forecast the inflation rate in China, but I can list some models I am going to do the empirical tests bellow.

**BP-ANN (Back Propagation – Artificial Neural Network)**

BP neural network has the ability to classify arbitrarily complex patterns and to do multi-dimensional function mapping. It can solve the Exclusive OR (XOR) and other problems that a simple perceptron can not handle. From a structural point of view, BP network has input layer, hidden layer and output layer. In essence, BP algorithm uses the square of network error as the objective function, and adopts gradient descent method to calculate the derivatives of the loss and thus minimize the objective/target function.

**SVM (Support Vector Machine)**

Support vector machines (SVM) are known as the classical supervised learning models with associated learning algorithms that analyze data and find one or several thresholds to classify these data. Given a set of training sets, every train set is labeled for a specific category, which is called the ground truth. An SVM training algorithm builds a model that assigns new examples to one category or the other. These labeled categories use gradient descent algorithm to optimize the model and make it a non-probabilistic binary linear classifier.

**Random Forest**

Semantically, random forest uses a random way to build a forest. There are many decision trees in the forest and trees are independent with one anther. After getting the forest, when a new input sample comes in, each decision tree in the forest will make its judgment separately, to see which direction the sample goes to (for the classification algorithm). The sample is judged as belonging to the category that has been most selected.

With the introduction of two randomizations, the random forest is less prone to over-fitting and has good anti-noise ability. It handles data of very high dimensions without the need for feature selection and well adapts to datasets: it handles both discrete and continuous data, and the dataset does not need to be normalized.

Random forest is a method of using multiple classification trees to discriminate and classify data. It classifies data and gives the importance score of each variable (gene), with wide application in classification and regression.

**K-Nearest Neighbor (KNN)**

The core idea of KNN algorithm is that if the majority of the k nearest neighbor samples in a feature space belong to a certain category, the sample also belongs to this category and has the characteristics of the sample in this category. The method determines the classification of the sample to be subdivided according to the category of the nearest one or several samples in determining the classification decision.

**Expected outcomes**

I expect that at least one model above can give more accurate forecast of China’s future inflation.

**References**

[1] 张国洪,曾永平等. 通货膨胀及紧缩与货币供应关系的实证分析[J]. 教师教育学报, 2005, 3(2):39-41.

[2] 朱连心. 本轮通货膨胀中的货币因素——基于月度数据的检验[J]. 当代经济, 2011(10):80-81.

[3] 张庆君, 姚树华. 汇率变动对通货膨胀影响的实证分析[J]. 统计与信息论坛, 2005, 20(5):75-78.

[4] 张强. 汇率和通货膨胀的关系研究[D]. 华中科技大学, 2009.

[5] 王维安, 贺聪. 房地产价格与通货膨胀预期[J]. 财经研究, 2005, 31(12):64-76.

[6] 王云清, 朱启贵, 谈正达. 中国房地产市场波动研究——基于贝叶斯估计的两部门DSGE模型[J]. 金融研究, 2013(3):101-113.

[7] Barth M J, Ramey V A. The Cost Channel of Monetary Transmission[J]. Nber Macroeconomics Annual, 2001, 16(Volume 16):199-240.

[8] 齐杨, 柳欣. 利率变动与通货膨胀[J]. 经济学动态, 2011(3):75-79.

[9] 刚猛, 陈金贤. 中国股票收益、通货膨胀与货币部门的角色分析[J]. 西安交通大学学报(社会科学版), 2004, 24(1):30-34.

[10] 刘金全, 王风云. 资产收益率与通货膨胀率关联性的实证分析[J]. 财经研究, 2004, 30(1):123-128.

[11] 李洪凯, 张佳菲, 罗幼强. 石油价格波动对我国物价水平的影响[J]. 统计与决策, 2006(6):81-83.

[12] 渠慎宁, 江贤武. 中国的经济增长与通货膨胀:基于产出缺口的实证解释[J]. 经济学动态, 2010(7):42-48.

[13] 王健, 田芬. 基于SVAR模型的货币政策与通货膨胀、GDP的动态关联分析[J]. 金田:励志, 2012(12).

[14] 肖曼君, Xia Rongyao. 中国的通货膨胀预测:基于ARIMA模型的实证分析[J]. 上海金融, 2008(8):38-42.

[15] Stock J H, Watson M W. Forecasting inflation ☆[J]. Journal of Monetary Economics, 2013, 44(2):293-335.

[16] Koop G, ‡ D K. FORECASTING INFLATION USING DYNAMIC MODEL AVERAGING \*[J]. International Economic Review, 2012, 53(3):867–886.

[17] Thakur G S M, Bhattacharyya R, Mondal S S. Artificial Neural Network Based Model for Forecasting of Inflation in India[J]. Fuzzy Information & Engineering, 2016, 8(1):87-100.

[18] Ülke V, Sahin A, Subasi A. A comparison of time series and machine learning models for inflation forecasting: empirical evidence from the USA[J]. Neural Computing & Applications, 2016:1-9.

[19] Mizrach B. Multivariate nearest-neighbour forecasts of ems exchange rates[J]. Journal of Applied Econometrics, 1992, 7(Supplement S1):S151–S163.

[20] Chen S, Härdle W K, Jeong K. Forecasting volatility with support vector machine-based GARCH model[J]. Journal of Forecasting, 2010, 29(4):406-433.

[21] "Machine Learning: What it is and why it matters". www.sas.com. Retrieved 2016-03-29.