Machine Learning applications in financial markets

B. Tech. Project Report

Submitted in partial fulfillment of the requirements
for the degree of

Bachelor of Technology

By

Prashant Pawar
Roll No: **06005007**

Under the guidance of

Prof. Saketh Nath

Department of Computer Science and Engineering
Indian Institute of Technology, Bombay
Mumbai

Acknowledgments

I would like to thank my guide, Professor Saketh Nath for his constant help, support and guidance.

Abstract

Stock market prediction with data mining techniques is one of the most important issues to be investigated. We intend to present a system that predicts the changes of stock trend by analyzing the influence of news articles.

Contents

Chapter 1

Introduction

Efficient Market Hypothesis

The **efficient-market hypothesis** (**EMH**) asserts that financial markets are "informational efficient", or that prices on traded assets (*e.g.,* stocks, bonds, or property) already reflect all known information, and instantly change to reflect new information. Therefore, according to theory, it is impossible to consistently outperform the market by using any information that the market already knows, except through luck. Information or *news* in the EMH is defined as anything that may affect prices that is unknowable in the present and thus appears randomly in the future. Stock market prediction brings with it the challenge of proving whether the financial market is predictable or not, since there has been no consensus on the validity of Efficient Market Hypothesis (EMH).

Stock market prediction has been an important issue in the field of finance, engineering and mathematics due to its potential financial gain. As a vast amount of capital is traded through the stock market, the stock-market is seen as a peak investment outlet. Researchers have strived for proving the predictability of the financial market. Henceforth, Stock Market prediction has always had a certain appeal for researchers. While numerous scientific attempts have been made, no method has been discovered to accurately predict stock price movement. Even with a lack of consistent prediction methods, there have been some mild successes.

Autoregressive and moving average are some of the famous stock trend prediction techniques which have dominated the time series prediction for several decays. With the help of data mining, several approaches using inductive learning for prediction have also been developed, such as k-nearest neighbour and neural network. However, their major weakness is that they rely heavily on structural data, in which they neglect the influence of non-quantifiable information such as news articles.

With the advent of faster computers and vast information over the Internet, stock markets have become more accessible to either strategic investors or the general public. Information from quarterly reports or breaking news stories can dramatically affect the share price of a security. As the Internet provides a primary source of event information which has a significant impact on stock markets, the techniques to extract and use information to support decision making have become a critical task. To predict the stock market accurately, various prediction algorithms and models have been proposed by many researchers in both academics and industry. In this paper, recent development in prediction algorithms and models will be introduced and their performance will be compared. In addition, for accurate stock market prediction, we investigate various global events and their issues on predicting stock markets

Chapter 2

Motivation

Consider the article published on 1st December on DAWN.com

Dubai debt crisis impact on dollar

The dollar slid against the euro and lingered near a 14-year trough against the yen, as dealers focused on Dubai, ahead of fresh US economic data, dealers said on Monday. In late morning trading here, the European single currency rose to $1.5076 from 1.4995 late in New York on Friday. Against the Japanese currency, the dollar fell to 86.34 yen from 86.52 yen late on Friday, but up from a nadir of 84.82 reached last week. “So what is the impact of the situation in Dubai? Well so far, nobody is really sure,” said Phil McHugh, senior dealer at Currencies Direct. “Given that both Abu Dhabi and the UAE central bank have quickly come forward to assure markets of their support for the beleaguered Emirates state, then the financial impact should be quite minor. He added that the safe-haven dollar would likely strengthen against rival currencies if the Dubai debt crisis worsens. “If things do begin to look a little dire, then expect the dollar to come back into focus as risk aversion trades re-emerge,” McHugh said. VTB Capital economist Andrew MacKinnon warned that Dubai could return to haunt global financial markets. “The threat of wider financial contagion from Dubai has proved limited so far--but Dubai is a reminder of sovereign credit risk, as well as problems related to commercial real estate loan exposure,” MacKinnon said. “We are not out of the woods yet and last week’s events have unsettled investor sentiment.”Meanwhile the euro firmed against the dollar ahead of an interest rate decision by the European Central Bank on Thursday. The bank is expected to maintain its key lending rate unchanged at 1 per cent, and it will likely raise its forecast for 2010 growth in the euro-sharing region, dealers said. In Asia, Bank of Japan Governor Masaaki Shirakawa said the bank was monitoring the impact of the stronger yen on businesses and stood ready to take necessary actions to stabilize the financial markets. The BoJ “pays due attention to the effects of the recent rapid appreciation of the yen on business sentiment,” Shirakawa said in a speech. “Stable moves in the foreign exchange market are desirable.” Money players were bracing themselves for a slew of US economic figures this week that could put fresh pressure on the dollar, dealers said. Factory and construction spending data will be out on Tuesday, followed by the Beige Book on economic conditions on Wednesday, and the key monthly unemployment report on Friday.—AFP

Stock variation on 1st December taken form KSE site

09:21:17|9206.21|6529.44|9706.48

09:25:07|9206.21|6529.44|9706.48

09:30:52|9206.21|6529.44|9706.48

09:35:22|9030.57|6414.4|9499.74

09:36:00|9013.48|6403.04|9484

09:40:29|9051.42|6425.48|9533.06

09:45:37|9023.16|6407.99|9504.12

09:50:07|8998.05|6391.3|9486.54

09:55:13|9027.12|6410.35|9512.3

10:00:20|9044.93|6422.13|9517.14

10:05:31|9052.15|6426.9|9539.73

10:10:01|9071.21|6439.37|9550.07

10:15:08|9064.31|6434.7|9538.14

10:20:18|9079.57|6445.13|9557.81

10:25:25|9089.59|6451.29|9566.03

10:30:31|9074.73|6441.55|9553.68

10:35:39|9084.59|6448.5|9565.16

10:40:46|9076.53|6443.19|9555.45

10:45:53|9088.15|6451.27|9568.67

10:51:00|9075.29|6443.04|9549.25

10:55:29|9068.61|6438.92|9540.09

11:00:36|9056.16|6430.71|9524.9

11:05:04|9065.35|6436.08|9538.4

11:10:11|9058.08|6430.41|9527.96

11:15:56|9055.39|6428.07|9527

11:20:40|9052.26|6426.08|9518.55

11:25:09|9044.8|6421.06|9507.12

11:30:17|9040.36|6418.68|9507.06

11:35:24|9045.36|6421.26|9516.97

11:40:31|9046.66|6421.77|9515.76

11:47:35|9038.6|6416.76|9503.95

11:50:08|9037.68|6416.08|9501.36

11:55:16|9036.22|6414.97|9500.03

11:59:44|9028.31|6409.39|9493.12

12:03:35|9024.47|6406.61|9487.17

12:07:28|9032.96|6412.04|9497.72

If we scan the document we can see, Dubai debt crisis word occurs in document and along with it other negative sentiments from the document and we can see that stock prices has fallen down during this period. Stock price went down from 9206.21 at 9.21am to 9032.96 at 12.07pm

This gives an idea of how news articles give a hint about stock market variations. So there is need to capture the impact of news articles for better prediction model.

Chapter 3

Literature Review

When predicting the future prices of Stock Market securities, there are several theories available. The first is Efficient Market Hypothesis (EMH) (Fama 1964). In EMH, it is assumed that the price of a security reflects all of the information available and that everyone has some degree of access to the information. Fama's theory further breaks EMH into three forms: Weak, Semi-Strong, and Strong. In Weak EMH, only historical information is embedded in the current price. The Semi-Strong form goes a step further by incorporating all historical and currently public information into the price. The Strong form includes historical, public, and private information, such as insider information, in the share price. From the tenets of EMH, it is believed that the market reacts instantaneously to any given news and that it is impossible to consistently outperform the market

A different perspective on prediction comes from Random Walk Theory (Malkiel 1973). In this theory, Stock Market prediction is believed to be impossible where prices are determined randomly and outperforming the market is infeasible. Random Walk Theory has similar theoretical underpinnings to Semi-Strong EMH where all public information is assumed to be available to everyone. However, Random Walk Theory declares that even with such information, future prediction is ineffective

It is from these theories that two distinct trading philosophies emerged; the fundamentalists and the technicians. In a fundamentalist trading philosophy, the price of a security can be determined through the nuts and bolts of financial numbers. These numbers are derived from the overall economy, the particular industry's sector, or most typically, from the company itself. Figures such as inflation, joblessness, industry return on equity (ROE), debt levels, and individual Price to Earnings (PE) ratios can all play a part in determining the price of a stock.

In contrast, technical analysis depends on historical and time-series data. These strategists believe that market timing is critical and opportunities can be found through the careful averaging of historical price and volume movements and comparing them against current prices. Technicians also believe that there are certain high/low psychological price barriers such as support and resistance levels where opportunities may exist. They further reason that price movements are not totally random, however, technical analysis is considered to be more of an art form rather than a science and is subject to interpretation.

Both fundamentalists and technicians have developed certain techniques to predict prices from financial news articles. In one model that tested the trading philosophies; LeBaron et. al. posited that much can be learned from a simulated stock market with simulated traders (LeBaron, Arthur et al. 1999). In their work, simulated traders mimicked human trading activity. Because of their artificial nature, the decisions made by these simulated traders can be dissected to identify key nuggets of information that would otherwise be difficult to obtain. The simulated traders were programmed to follow a rule hierarchy when responding to changes in the market; in this case it was the introduction of relevant news articles and/or numeric data updates. Each simulated trader was then varied on the timing between the point of receiving the information and reacting to it. The results were startling and found that the length of reaction time dictated a preference of trading philosophy. Simulated traders that acted quickly formed technical strategies, while traders that possessed a longer waiting period formed fundamental strategies (LeBaron, Arthur et al. 1999). It is believed that the technicians capitalized on the time lag by acting on information before the rest of the traders, which lent this research to support a weak ability to forecast the market for a brief period of time.

In similar research on real stock data and financial news articles, Gidofalvi gathered over 5,000 financial news articles concerning 12 stocks, and identified this brief duration of time to be a period of twenty minutes before and twenty minutes after a financial news article was released (Gidofalvi 2001). Within this period of time, Gidofalvi demonstrated that there exists a weak ability to predict the direction of a security before the market corrects itself to equilibrium. One reason for the weak ability to forecast is because financial news articles are typically reprinted throughout the various news wire services. Gidofalvi posits that a stronger predictive ability may exist in isolating the first release of an article. Using this twenty minute window of opportunity and an automated textual news parsing system, the possibility exists to capitalize on stock price movements before human traders can act.

In similar research by Robert P. Schumaker, they picked research period of Oct. 26 to Nov. 28, 2005, to gather news articles and stock quotes from S&P 500. They gathered around 9211 financial news articles and 10,259,042 stock quotes over the five week period. They analysed the news articles using the three textual representations and retained only those terms that appeared three or more times in an article, which
results in a differing number of articles. The filtering process resulted in the following breakdown:

—Bag of words used 4,296 terms from 2,839 articles
—Noun phrases used 5,283 terms from 2,849 articles
—Named entities used 2,856 terms from 2,620 articles

Then this was processed by Support Vector Machine derivative, using Sequential Minimal Optimization in a form of regression, which can handle discrete number analysis.
They chose two evaluation metrics; closeness and directional accuracy.
They had three different models, first model (M1) uses only extracted article terms for its prediction. While no baseline stock price exists within this model. Second model (M2) uses extracted article terms and the stock price at the time the article was released. Third model (M3) uses extracted terms and a regressed estimate of the +20 minute stock price. This model may lead to better predictive results should the article terms have no impact on the movement of the stock price. All three models rely on using article terms in their prediction. SVM learns what terms lead to share price changes and adjust their weights depending on the severity of price changes

Results for this research

Closeness Results

Mean Square Error	Regress	M1	M2	M3
Bag of words	0.07279	930.87	0.04422	0.12605
Noun Phrases	0.07279	863.50	0.04887	0.17944
Named Entities	0.07065	741.83	0.03407	0.07711
Average	0.07212	848.15	0.04261	0.12893

Directional Accuracy Result

Directional Accuracy	Regress	M1	M2	M3
Bag of Words	54.8%	52.4%	57.0%	57.0%
Noun Phrases	54.8%	56.4%	58.0%	56.9%
Named Entities	54.2%	55.0%	56.4%	56.7%
Average	54.6%	54.6%	57.1%	56.9%

From looking at the average results in Table, Model M2 which used both article terms and the stock price at the time of article release, had the lowest MSE score (0.04261) of any of the models (p-values < 0.05). This result signifies that Model M2's predictions were closer to the actual +20 minute stock price than any of the other models including linear regression (regress). Looking deeper into the results, we find that Model M2 performed better than regress in each of the three textual representations.

For model M2, weighting scheme that SVM assigned to the training variables shows that, stock price at the time of article release was given a weight of 0.9997 by SVM, while the article terms had a combined weight of 0.0003. As the results have shown that model M2 which uses news articles unlike regress performs better than regress. So the superficially light weight of 0.0003 also plays an important role in stock value prediction, this shows the importance of news articles in stock value prediction problem, this gives us a motivation to incorporate the impact of news articles diligently for better performance.

3.1 Textual Representation

There are a variety of methods available to analyze financial news articles. One of the simplest methods is to tokenize and use each word in the document. While this human friendly approach may help users to understand the syntactic structure of the document, machine learning techniques do not require such structural markings. This technique also assigns importance to determiners and prepositions which may not contribute much to the gist of the article. One method of circumventing these problems is a Bag of Words approach. In this approach, a list of semantically empty stop-words are removed from the article (e.g.; the, a, and for). The remaining terms are then used as the textual representation. The Bag of Words approach has been used as the de facto standard of financial article research primarily because of its simple nature and ease of use. Building upon the Bag of Words approach, another tactic is to use certain parts of speech as features. Noun Phrasing is accomplished through the use of a syntax where parts of speech (i.e., nouns) are identified through the aid of a lexicon and aggregated using syntactic rules on the surrounding parts of speech, forming noun phrases. A third method of article representation is Named Entities. This technique builds upon Noun Phrases by using a semantic lexical hierarchy where nouns and noun phrases can be classified as a person, organization, or location. This hierarchy operates by analyzing

the synonyms of each noun and generalizing their lexical profile across the rest of the noun phrase. Named Entities in effect provide for a more abstract representation than Bag of Words or Noun Phrases.

As the result from Robert P. Schumaker experiment shows that on an average named entity approach of textual representation performs better. So which textual representation should be used is also an important question to ask. A semantic analysis of texts can also be used as an efficient method for representation.

3.2 Machine Learning Algorithms

Like textual representation, there are also a variety of machine learning algorithms available. Almost all techniques start off with a technical analysis of historical security data by selecting a recent period of time and performing linear regression analysis to determine the price trend of the security. From there, a Bag of Words analysis is used to determine the textual keywords. These outcomes are then classified into stock movement prediction classes such as up, down, and unchanged. Much research has been done to investigate the various techniques that can lead to stock price classification. Following Table illustrates Stock Market prediction taxonomy of the various machine learning techniques.

Algorithm	Classification	Source Material	Examples
Genetic Algorithm	2 categories	Undisclosed number of chatroom postings	Thomas & Sycara, 2002
Naïve Bayesian	3 categories	38,469 articles	Lavrenko et al. 2000
SVM	3 Categories	6602 articles	Mittermayer, 2004

From above Table, several items become readily noticeable. The first of which is that a variety of techniques have been used. The second is that almost all instances commonly classify predicted stock movements into a set of classification categories, not a discrete price prediction. Lastly, not all of the studies were conducted on financial news articles, although a majority was. The first technique of interest is the Genetic Algorithm. In this study, discussion boards were used as a source of independently generated financial news (Thomas and Sycara 2002). In their approach, Thomas and Sycara attempted to classify stock prices using the number of postings and number of words posted about an article on a daily basis. It was found that positive share price movement was correlated to stocks with more than 10,000 posts. However, discussion board postings are quite susceptible to bias and noise. Another machine learning technique, Naïve Bayesian, represents each article as a weighted vector of keywords. Phrase co-occurrence and price directionality is learned from the article which leads to a trained classification system. One such problem with this style of machine learning is from a company mentioned in passing. An article may focus its attention on some other event and superficially reference a particular security. These types of problems can cloud the results of training by unintentionally attaching weight to a casually-mentioned security. One of the more interesting machine learners is Support Vector Machines (SVM). In the work of Fung et. al., regression analysis of technical data is used to identify price trends while SVM analysis of textual news articles is used to perform a binary classification in two predefined categories; stock price rise and drop (Fung, Yu et al. 2002). In cases where conflicting SVM classification ensues, such as both rise and drop classifiers are determined to be positive, the system returns a 'no recommendation' category. From their research using 350,000 financial news articles and a simulated Buy-Hold strategy based upon their SVM classifications, they showed that their technique of SVM classification was mildly profitable. Mittermayer also used SVM is his research to find an optimal profit trading engine. While relying on a three tier classification system, this research focused on empirically establishing trading limits. It was found that profits can be maximized by buying or shorting stocks and taking profit on them at 1% up movement or 3% down movement. This method slightly beat random trading by yielding a 0.2% average return.

Chapter 4

Data collection

Data collection was one of the biggest tasks which we had to face, we first look out for BSE, but it gives only opening & closing price. We scanned through almost all the indices out there in the world to see if anyone has archive for the intra-day variations.

These are all the indices which we tried out for

http://www.tdd.lt/slnews/Stock_Exchanges/Stock.Exchanges.htm

African stock exchanges, Asian stock exchanges, European stock exchanges, Middle Eastern Stock Exchanges, North American Stock Exchanges, South American Stock Exchanges

Then we even look for WRDS - Wharton Research Data Service which gives data for academic purposes but unfortunately IIT doesn't have account on this site.

Finally we found Karachi stock exchange gives intra-day changes for that day, so we had collect data from KSE for last 30 days but this also did not work out as the news articles for the same did not have the proper time tag and availability was also less.

Then we came across Yahoo Finance website, it has the intraday 5 min data available for that day, but collecting it was very difficult, we manually collected the data from yahoo finance site along with the news articles from Reuters, Indian Express etc.

Chapter 5

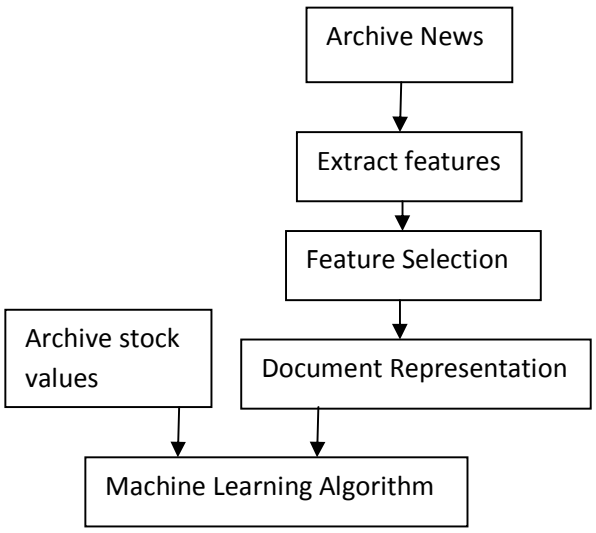
Results

So there are two important questions which need to be addressed:

1. How to catch the impact of news articles on stock market?
2. How to efficiently retrieve available information from news articles?

Many research works has tried to address above questions, research till now has used the regression methods to catch the impact, namely simple regression on stock values, regression on news articles along with base value of stock, regression on stock value along with regression on news articles.
We intend to use multimodal agent approach which is not been used in the field before.

Till now research has used different text representation methods namely bag of words, named entities, noun phrases, we intend to use bag of words approach to start off our work. Then we intend to sentiment analysis of document which has not been tried till now in this field.



We followed this and performed experiments using Rapid-miner.

$F(n) = S_{t+20}$

This model captures the effect of only news articles on stock market index, this model may not perform well as it assumes stock market is driven solely by news articles, as we all know this is not the case. This model has the benefit of giving directions which are important for trading strategies. The linear regression and SVM regression was performed and the following results were found.

Measure	Linear Regression	SVM Regression
Root mean squared error	11,752.259 +/- 0.000	11,752.299 +/- 0.000
Relative error	3.82% +/- 6.96%	3.81% +/- 6.96%
Normalized absolute error	0.798	0.797
Root relative squared error	1.001	1.001

$F(n, S_t) = S_{t+20}$

This model captures the effect of news articles and the base stock market value to get the value after 20 minutes (as news article has impact for 20 minutes). This model will perform better as it also includes the base value. This model has the benefit of giving directions as well as the quantification of how much value has changed. The linear regression and SVM regression was performed and the following results were found.

Measure	Linear Regression	SVM Regression
Root mean squared error	11,752.375 +/- 0.000	11,826.275 +/- 0.000
Relative error	3.80% +/- 6.96%	3.88% +/- 7.00%
Normalized absolute error	0.796	0.804
Root relative squared error	1.001	1.001

Chapter 6

Correlation between stock markets

With trend of globalization in international financial market, more intimate of capital flows was formed day by day in capital market. Thus, it is important to understand the relationship of individual country in international stock markets. As we have seen the stock market index of one country has impact on others and vice versa. So we tried to learn how a one stock market index has the impact on other and how to capture this correlation. Different techniques like Co-integration Test, Granger Causality etc were used to identify whether stock markets exist long-term relationship.

The dependence of the correlation directly depends on the level of prevailing uncertainty, which is measured in terms of volatilities and other potential risk factors. Model was built to investigate how time varying correlation is affected by internal volatilities (i.e., volatility terms included in the correlation definition), external volatilities (i.e., volatility terms not included in the correlation definition), and other factors such as market trend [10]. Research till now has shown that correlations are significantly higher when world markets are down trending. Varying correlations between stock market returns primarily explained by internal national market volatilities and external world market volatilities. Moreover, in terms of economic significance, we find that large increases in volatility can substantially change correlations. Down trends in world markets are significant at times but have a relatively weaker relationship than volatility to correlations between stock market returns.

The conditional correlation between series ut and vt is given by,

$$\rho_t(u,v) = \frac{\operatorname{cov}_t(u,v)}{\sqrt{\operatorname{var}_t(u)\operatorname{var}_t(v)}}$$

From above equation we see that, if correlation is time invariant then time-varying covariance must change in a fixed proportion to the product of the time-varying standard deviations Consequently, it may be difficult to infer on the basis of the covariance whether the dependence per se between the series is time-varying or due simply to the fixed relation between the volatilities and covariance determined by the time invariant correlation.
Different arch models were used to capture the correlation.

Long term relationship between US, China and Japan stock market were studied. Different methods were used to see if one stock market index has any dependence on other.

Causality test- Suppose two time series X, Y, when predicting X, if adding past value of Y to information set would get more accurate prediction result of X, then it indicates “Y is the cause of X”, alternatively, X might be the cause of Y, too. Causality test was performed to get relationship.

The results found due to this, suggested that stock market indices present severe volatility. Specially Us stock market has more dependence than China and Japan stock market.

Chapter 7

Conclusion

We have successfully performed the experiments which captures the effect of news articles on stock market. The results have shown that there is stronger information quotient in news articles which is essential for the stock market index prediction.

We have also gone through the research which tries to capture the correlation between two different stock markets. But unfortunately due to time constraint we did not able to perform experiment for the same.

Chapter 8

Future work

A lot of scope for research still exists as this is a very hot topic and a little improvement is also very valuable. So, one can pose this problem as "multi-modal regression" and look for benefits. One can also include sentiment analysis in our text analysis to improve the prediction accuracy.

References

1. Textual analysis of stock market prediction using financial news articles by Robert P. Schumaker and Hsinchun Chen February 2009
2. News Sensitive Stock Trend Prediction by Gabriel Pui Cheong Fung, Jeffrey Xu Yu, and Wai Lam 2002
3. Machine learning techniques and use of event information for stock market prediction: a survey and evaluation by Paul D. Yoo, Maria H. Kim, Tony Jan 2007
4. A Multi agent approach to Q-learning for daily stock trading by Jae Won Lee, Jonghun Park, *Member, IEEE*, Jangmin O, Jongwoo Lee, and Euyseok Hong 2007
5. Intelligent Stock Trading System based on SVM Algorithm and Oscillation Box Prediction by Qinghua Wen, Zehong Yang, Yixu Song, Peifa Jia
6. Forecasting Intraday Stock Price Trends with Text Mining Techniques by Mittermayer, *University of Bern, Institute of Information Systems 2004*
7. Document Clustering for Event Identification and Trend Analysis in Market News by Lipika Dey, Anuj Mahajan and SK. Mirajul Haque
8. Enabling Sophisticated Financial Text Mining by Calum Robertson 2009
9. An Approach to Text Mining using Information Extraction by Haralampos Karanikas, Christos Tjortjis and Babis Theodoulidis
10. What What Drives Correlation Between Stock Market Returns? International Evidence by Johan Knif, James Kolari, Seppo Pynnönen
11. The Association of Stock Index among the market of China, US., and Japan by Meng-Long Shih, National Taitung University