Project Proposal

Submitted By - Anuroopa Tilavalli Shettar & Divya Singh

We are presenting two proposals -
1. Crime Prediction in Boston
2. Stock Prediction for Tesla using Sentiment Analysis

Idea 1: Crime Prediction in Boston

Problem: According to NeighbourhoodScout's analysis, the crime rate in Boston is considerably higher than the national average across all communities in America from the largest to the smallest. The chance of becoming a victim of either violent or property crime in Boston is 1 in 34. Based on FBI crime data, Boston is not one of the safest communities in America. Relative to Massachusetts, Boston has a crime rate that is higher than 97% of the state's cities and towns of all sizes.

Violent offenses tracked included murder and non-negligent manslaughter, armed robbery, and aggravated assault, including assault with a deadly weapon. Boston's rate for property crime is 23 per one thousand population. Property crimes are motor vehicle theft, arson, larceny, and burglary.

BOSTON VIOLENT CRIMES				POPULATION: 685,094
	MURDER	RAPE	ROBBERY	ASSAULT
Report Total	57	314	1,505	2,940
Rate per 1,000	0.08	0.46	2.20	4.29

Solution: Our solution is to provide crime prediction. Although, predicting the crime will not prevent it from happening, but to some extent it will provide security to crime sensitive areas. Our job is to find the crime hotspots using classification methods based on location and time.

Input: We would be using input datasets from kaggle which consist of crime incident reports provided by Boston Police Department (BPD) to document the initial details surrounding an incident to which BPD officers respond. This is a dataset containing records from the new crime incident report system, which includes a reduced set of fields focused on capturing the type of incident as well as when and where it occurred.

Dataset: Our dataset has 17 features. We will preprocess the data to filter out records with missing entries. A feature will be added for location based crime percentage using heuristic evaluation which be later used as a standard to test the algorithm output against. We will divide the dataset into 80:20 ratio where 80% of data will be used for training and rest 20% for testing.

Data Reduction: Out of 17 entries we will filter out columns which are not relevant to the study.

Output: Boston area map visualization of with highlighted areas the most crime prone.

Algorithms
What algorithms do you use?
- Naive Bayes (Divya)
- Decision Trees (Divya)
- Random Forest (Anuroopa)
- K Nearest Neighbour (Anuroopa) *[handwritten: ← this is simple, want something new instead or in addition]*

Why are these algorithms appropriate?
All these algo are appropriate because we are classifying areas in the city which are crime hotspots or not. We have labeled training data set so it will fall under supervised learning. And since our classification is binary.

How are these algorithms typically used, and how are you using them?
Decision Trees is a non-parametric supervised learning method that predicts the value of a target variable by learning simple decision rules inferred from the data features.
Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features.
[handwritten margin note: more detail] A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.
kNN is a classifier that makes a prediction based on the majority vote of the k nearest samples on the feature vector space.

We are using these algorithms to classify the areas in boston in to two categories crime hotspot and not a crime hotspot based on the crime rate.

Have other people use similar algorithms to solve your problem before?
Examining Deep Learning Architectures for Crime Classification and Prediction By Panagiotis Stalidis, Theodoros Semertzidis and Petros Daras [3]
kNN Classifier and Naive Bayes Classifier for Crime Prediction in San Francisco by Noora Abdulrahman and Wala Abedalkhader [2]

Results

What results do you expect to show?
We will show the results through visualization by highlighting the crime hotspots in the Boston area.

What comparisons will you do?
We will compare the predictions based on heuristic evaluation methods.

Are there risks for not getting all the results?
Based on previous research we have seen that these are effective algorithm to get quite accurate results.

If so, what will you do about it? *[handwritten: need exclusive analysis]*

We will do empirical analysis to to improve our algorithms

Idea 1: Stock Prediction for Tesla using Sentiment Analysis

Problem: According to TheStreet, when it comes time to make investment decisions, it's a good idea to be guided by more than just gut instincts. If used effectively, fundamental analysis is one of the most useful ways to determine whether a company is a good investment choice. The stock (also capital stock) of a corporation is all of the shares into which ownership of the corporation is divided [4]. The term stock trader typically refers to someone who frequently buys and sells stocks to capitalize on daily price fluctuations [5]

A forecast is used to have a good idea of what is likely to take place. Determining the future value of a company's stock or other financial instrument traded on exchange either by fundamental or technical analysis is called stock market predictions. The advantage of predicting the stock market is it helps you to invest wisely to make good profits. The advantage of predicting the stock market is to earn good profit. [6] Additionally, the public sentiment about the stock also plays a major role. It's about what hear from people and trusting what most people are doing.

Solution: Our solution is to provide stock prediction for Tesla based on computational analysis and public sentiment. This will help the user to get predictions about the Tesla stock constructed using technical analysis along with consensus opinion.

Input: We would be using input datasets from Intrino which provides detailed financial data through rest services. The dataset will consist of stock, historical and other standardized financial details. The dataset can be extracted using parameters such as return dates, exchange, and identifier as Tesla.

Dataset: Our dataset will have multiple financial features such as Revenue, Gross Profit, Total Debt, current Ratio, return on equity, market capitalization etc. We will preprocess the data to transform it into a suitable data structure as input for the algorithms. We will divide the dataset into 80:20 ratio where 80% of data will be used for training and rest 20% for testing.

Output: A prediction for the stock price of Tesla based on sentiment analysis.

Algorithms
What algorithms do you use?
- Linear Regression (Divya)
- Random Forest (Divya + Anuroopa)
- Support Vector Machines (Anuroopa)

Why are these algorithms appropriate?
Both of these algorithms are appropriate because both of them are used for predictive analysis, which is the goal of our project.

All these algo are appropriate because we are classifying areas in the city which are crime hotspots or not. We have labeled training data set so it will fall under supervised learning. And since our classification is binary.

How are these algorithms typically used, and how are you using them?

Linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable. This is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding out cause and effect relationship between variables.

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.

We are using these algorithms to predict the stock price providing financial variables and sentiment analysis as input through predictive analysis.

Have other people use similar algorithms to solve your problem before?

Market Trend Prediction using Sentiment Analysis: Lessons Learned and Paths Forward By Andrius Mudinas, Dell Zhang, Mark Levene [7]

Stock Prediction Using Twitter Sentiment Analysis By Anshul Mittal and Arpit Goel [8]

A Sentiment Analysis Approach to Predicting Stock Returns By Tom Yuz [9]

Results

What results do you expect to show?

A prediction for the stock price of Tesla based on sentiment analysis in the form of accuracy percentage.

What comparisons will you do?

We will compare the predictions based on real time stock values.

Are there risks for not getting all the results?

Based on previous research we have seen that these are effective algorithm to get quite accurate results.

If so, what will you do about it?

We will do empirical analysis to to improve our algorithms

References

1. https://www.neighborhoodscout.com/ma/boston/crime
2. http://aircconline.com/ijdms/V9N4/9417ijdms01.pdf
3. https://arxiv.org/pdf/1812.00602.pdf
4. https://en.wikipedia.org/wiki/Stock
5. https://www.nerdwallet.com/blog/investing/stock-trading-how-to-begin/
6. https://www.quora.com/What-are-the-benefits-of-stock-market-predictions
7. http://sentic.net/wisdom2018mudinas.pdf
8. http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf
9. https://medium.com/@tomyuz/a-sentiment-analysis-approach-to-predicting-stock-returns-d5ca8b75a42