**RESUME**

* My name is Hazel Bing. I graduated from Columbia University, master in Operations Research. And I am looking for a full-time data scientist position.
* In my mind, data scientists are the ones who should have the ability to tell a story from the data, and tell people what are not easily observed from the data without analytics. In this way, they add value and create more revenue to the business.
* For me, I have experience applied machine learning techniques solving real-life problems using Python and SQL, like building a recommendation system, some predictive models, etc; besides, I also have some experience dealing with large datasets. And I really enjoy doing this and hope to go further in this way. I love extracting information from data and find some valuable insights.

* And to pratice my hands-on experience, I did some data analysis projects, including customer churn prediction, moview rating prediction and recommendation and an NLP project.
* For the internship experience, last summer, I worked for a fin-tech startup, ForwardLane, where...
* And recently at OmniMarkets, I helped develop some challenger model in option pricing using machine learning algorithms. SVR and GBDT.

**--Internships**

1. OmniMarkets
   * **Big picture**
     + The traditional option pricing method is using Black-Scholes formula, while the assumptions of the method are often violated in reality. So we want to train an algorithm to learn from the history data in the market and make prediction without making any assumptions about the relationships between input variables.

* Background -- Methodology -- Data and Results -- Conclusion
* Applied SVR and Gradient Boosting Decision Tree in option pricing prediction using data from 2002 to 2016
* Tuned model parameters by grid search and improved pricing accuracy over 40% compared to traditional BS Model
* Organized weekly panel discussion for peer review, performed presentations to risk and management teams.

Background

Challenger model is to validate the results and evaluate the performance of the model over time.

The Black-Scholes approach to option pricing is one of the most important ideas in finance.

* Assumptions
* Inputs
* Output
* Shortcoming

Related work

In recent years, nonparametric methods are used for prediction. Because they are suitable to adjust changing behavior of the derivative securities as they can be trained from time to time.

Since 1994, neural network are used in option pricing and it performed fairly well.

Support Vector Machine is a powerful tool and it has been widely used for classification, pattern recognition and nonlinear function estimation. And it's found that when it's used for regression, it can provide much better results than neural network.

Gradient Boosting Decision Tree is also a method proved to be powerful in option pricing.

Methodology

Data preparation

Performance measurement

Conclusion

SVR can further be improved by changing the input variables in different conditions. for example, deep-in-the-money option, at-the-money option, out-of-the-money option.

Gradient boosting decision tree

grid search

Background

Challenger model is to validate the results and evaluate the performance of the model over time.

The Black-Scholes approach to option pricing is one of the most important ideas in finance.

Assumptions:

1. The price of a stock follows a geometric Brownian motion with constant volatility.

2. efficient markets. stocks move like random walk. At any given moment, the price of the underlying stock can go up or go down with the same probability.

3. No dividends.

4. constant and known interest rates.

5. No transaction costs and commissions.

Inputs:

Current stock price

Option strike price

time to expiration (denoted as a percent of year)

risk-free interest rates

Output:

Option price

Shortcoming:

In practice, assume the BS formula is correct and to solve the volatility by inverting the formula. However, BS implies that the volatilities obtained from options on the same stock are constant across different strikes K and maturities T. And prices in the markets do not exactly come from BSM.

So it is possible to train an algorithm to learn from the history data in the market and make prediction without making any assumptions about the relationships between input variables.

And these machine learning models are nonparametric

-----------

Related work

In recent years, nonparametric methods are used for prediction. Because they are suitable to adjust changing behavior of the derivative securities as they can be trained from time to time.

Since 1994, neural network are used in option pricing and it performed fairly well.

SVM: Support Vector Machine is a powerful tool and it has been widely used for classification, pattern recognition and nonlinear function estimation. And it's found that when it's used for regression, it can provide much better results than neural network.

GBDT: Gradient Boosting Decision Tree is also a method proved to be powerful in option pricing.

----------

Methodology

SVM

use a kernal to transform the data to one dimension, and make regression

1. pro

\* fewer hyperparameters with robust result

\* no assumptions of the non-linear functional transformation

2. con

\* time consuming

GBDT

1. Regression Decision Tree

builds regression in the form of a tree structure.

breaks down a dataset into smaller and smaller subsets.

the result is a tree with decision nodes and leaf nodes.

tree regression models calculate the relative importance of predictors, and the relative importance of predictors can be computed by sum of squared error.

it's like a complicated linear regression at each leaf node, and there is a prediction score in each leaf.

1. Gradient boosting

assume the tree is not a perfect model, so every time, each successive tree uses the residuals of the previous tree. in this way, it has strong predictive power.

shrinkage, give a weight to each tree

It is a ensemble of many trees.

through control the learning rate, number of estimations, max\_depth and max\_features, etc.

1. pros

\* fast

\* doesn't require czreful normalization of features to perform well.

1. cons

\* difficult to interpret

\* require the contraint of learning rate

Data preparation

1. AAPL, from 2002 to 2010.

2. Inputs: underlying price , strike price, time to expiration, volatility, risk-free interest rate, Black-Scholes option price.

3. remove data that have trading volume less than 100.

4. remove data that have zero trading volume

5. all options with less than 6 days or more than 260 days to expiration are eliminated to avoid extreme option prices that are observed due to potential liquidity problem.

6. after the data cleaning, there are 215393

7. split the data into training data set 70% and testing data set 30%.

Performance measurement

Explained Variance

MSE

Conclusion

SVR can further be improved by changing the input variables in different conditions. for example, deep-in-the-money option, at-the-money option, out-of-the-money option.

1. ForwardLane
   * **Big picture**
     + ForwardLane is a fin-tech startup based in New York, it is building a robo-advisor platform which can provide real-time investment solutions to the investors. I helped develop a recommendation model for optimizing the portfolios by applying ***clustering algorithm*** and ***Monte Carlo simulation***. The goal of this model is to give online clients customized recommendations to improve their current portfolio. And the language I use are mainly **Python** and **SQL**.
   * Using current portfolio's components and the investor's risk appetite as input, the model will automatically add or delete some elements in the portfolio and then allocate the proportion of the assets to satisfy clients' need.
   * **For example**, if you are a user of this platform, after you log in and enter into the conversation interface, it will ask your current asset allocation and your risk appetite. Then, it will analyze your portfolio structure and optimize it.

* **How to optimize? First**, it will determine which financial products should be in your portfolio, based on correlation analysis, some elements will be deleted, or it will select some other well-performing products from our database. And after choosing the right products, **the next step** is to adjust the allocation based on **efficient frontier** and your risk appetite.

* In this process, how to select products for adding them to your portfolio? I implemented clustering algorithm.

* ***Why* choose cluster?** For example, for the ETF database, there are thousands of products and it is impossible to try one by one because it is quite time-consuming. So I will firstly cluster all the ETFs to different clusters based on their **similarity** *of risk and return profile* using k-means cluster algorithm. In this way I can smaller the ETF universe.

* **How to cluster?** 
  + **K-means**: I use k means clustering algorithm. The goal of this algorithm is to find groups in the data, and the number of groups is the variable k. Data points are clustered based on feature similarity. The output of this method is the central point of the K clusters, which can be used to label other data. And each data point will be assigned to a single cluster.
  + Finally, The ETFs are grouped into 11 clusters.
  + **How to choose k?** In general, there is no method for determining exact value of *K*, but an accurate estimate can be obtained using the following techniques. A commonly used method is called Elbow method. It will first compute the sum of squared error (SSE) for some values of k. The SSE is defined as the sum of the squared distance between each member of the cluater and its centroid. Then we plot k against the SSE, we will see that *the error decreases as k gets larger*; this is because when the number of clusters increases, they should be smaller, so distortion is also smaller. The idea of the elbow method is to choose the k at which the SSE decreases abruptly. It means most of the variance can be explained when at this k.

* Optimization part

Important

**How to draw efficient frontier?** I ran thousands of times **Monte Carlo simulation** in Python to simulate as much as possible combinations of this portfolio, and every possible combination of assets plotted on graph. Plots return and risk(std. deviation). Optimal portfolio lies on the efficient frontier curve.

* This is how this recommendation model works.

* Monte Carlo simulation: use repeated random sampling to obtain numerical results
* How do you do the classification?

For each ETF, we calculate the **efficient frontier** to go back for one year historical data, we calculate the volatility, which we consider risk for this period, we calculate the expected return for the last year period, and then you have the risk and return profile, the next what we do is o draw a coordinate axis, the x axis is the risk and y is your return, for each combination, you have a point in the coordinate, the similar risk and return profile as one cluster.

* Evaluated the portfolio's performance using simulating the efficient frontier and Raise Partner's API
  + <https://raisepartner.com/?lang=en>
  + Compute ex-ante portfolio analytics and multi-classification contributions to understand the sources of risk and performance in your portfolio.
  + Define optimal and robust portfolios while meeting regulatory and business constraints
* **Efficient frontier background theory**

NLP project:

The goal is categorize financial news and its related question and answer;

If we have news come in today, we can category these news based on the classification we did.

News is json file text data. Use python NLTK package to clean the text data. The detail process is:

LDA:

5.1. What is LDA? 
Latent Dirichlet allocation (LDA) is a generative statistical model. 
LDA is an example of a topic model. 
In LDA, each document is assumed to be characterized by a particular set of topics. 
If observations are words collected into documents, LDA result means that: 
Each document is a mixture of a small number of topics. 
o 
Each word's creation is attributable to one of the document's topics. 
o 

Each news is assumed to be characterized by a particular set of topics

If observations are words collected into documents, LDA means that

Each news is a mixture of a small number of topics

Each word's creations is attributable to one of document's topics

LDA could tags each document with topics

Steps:

Tokenizing and stemming:

Load stopword and drop stopword

TF-IDF:

We need a weighting factor to reflect the importance of a word in collection or corpus

Term frequency and inverse document frequency

Choose N-gram: range

Get tfidf matrix

Document similarity:

Use tf-idf matirx to calculate similarity

Cosine similarity

K-means cluster:

Use PCA and keep the first and second component. Plot the 1 and 2 to see the cluster result

MVO:

Introduction. The fundamental goal of portfolio theory is to optimally allocate your investments between different assets. Mean variance optimization (MVO) is a quantitative tool which will allow you to make this allocation by considering the trade-off between risk and return.

An efficient portfolio is either a portfolio that offers the highest expected return for a given level of risk, or one with the lowest level of risk for a given expected return. The line that connects all these efficient portfolios is the efficient frontier.

B-L Model:

The Black-Litterman model is a model used to estimate inputs for portfolio optimization. It mixes different types of estimates, some based on historical data, others based on equilibrium conditions to arrive at updated estimates

The beauty of this model is that one can blend a variety of views specified in different ways, absolute or relative, with a given prior estimate to generate a new and updated posterior estimate which includes all the views.

**--Projects**

problem -- variables -- features -- algorithm -- strategy -- accuracy

1. Kaggle Toxic comment classification
2. IMDB Movie Data Analysis
3. User Churn Prediction