Options Pricing Group Project

For this project you will be examining European call option pricing data on the S&P 500. Recall, a European call option gives the holder the right (but not the obligation) to purchase an asset at a given time for a given price. Valuing such an option is tricky because it depends on the future value of the underlying asset.

The Black-Scholes option pricing formula provides an approach for valuing such options. Let K denote the strike price, i.e., the price one must pay to purchase the asset, and τ (tau) the time until expiration of the option. Suppose that the asset in question is currently trading at S, and has "volitility" (i.e., risk or standard deviation) of σ. Finally, suppose that the annual risk-free interest rate is r. Then the Black-Scholes formula states

$$C\_pred = S\Phi(d_1) - Ke^{-r\tau}\Phi(d_2),$$

where C_pred is the predicted option value, $d_1 = \frac{\log\left(\frac{S}{K}\right)+(r+\sigma^2)\tau}{\sigma\sqrt{\tau}}$, $d_2 = d_1 - \sigma\sqrt{\tau}$ and $\Phi(x)$ represents the probability that a standard normal random variable will take on a value less than or equal to x.

The 1997 Nobel Prize in Economics was awarded for the Black-Scholes formula because it works remarkably well in practice. However, in this project we are going to attempt to build statistical/ML models to perform the same task. *In this project, you should pretend that you don't know the Black-Scholes formula when building your machine learning models*.

You will find two data sets on Blackboard: option_train.csv and option_test_nolabel.csv. The training data set has information on 5,000 separate options. In particular, for each option we have recorded

- Value (C): Current option value
- S: Current asset value
- K: Strike price of option
- r: Annual interest rate
- tau: Time to maturity (in years)
- BS: The Black-Scholes formula was applied to this data (using some σ) to get C_pred. and If an option has C_pred – C > 0, i.e., the prediction over estimated the option value, we associate that option by (Over); otherwise, we associate that option with (Under).

The test data set is similar except it has only 500 options and is missing the Value and BS variables. You can safely assume that the test data is of good quality, but you should check for missing and erroneous entries in the training data.

The core idea of the project is to use the training data to build statistical/ML models with

1) Value as the response (i.e., a regression problem) and then
2) BS as the response (i.e., a classification problem).

The other four variables will be used as the predictors. You will explore all of the regression (for Value) and classification (for BS) methods we have discussed in the course on the training data (you may also use methods we have not discussed, but this is not required). Ultimately you will select what you consider to be the most accurate approach and use it to make predictions for C and BS on the 500 options in the test data set. You will submit these two sets of predictions. I will compare these predictions in comparison to the actual Value and BS results on the test options (which I have), in terms of **out-of-sample R squared** and **classification error**, respectively.

For BS you must submit a column of 1's and 0's (not words or probabilities!) with 1 corresponding to a prediction of "Over" and 0 a prediction of "Under".

You submit your predictions for Value and BS in csv file with two columns (with Value and BS as the column names). For example, group x should submit **group_x_prediction.csv**. Please follow this naming convention. See the sample submission attached (**group_0_prediction.csv**).

Grading

The project will be graded out of **13 points.** 7 points will be allocated to the project report and 6 points will be allocated to the presentation stage. **Every member of the group should speak in the presentation.** And the video of the speakers should appear in the recording.

Section		**Points**	**Comments**
Project Report	Write Up	3	See the next page for further instructions.
	Value Prediction	2	It is easy to get 90%. So I will allocate 0 point for <90%, 1 for between 90% and 94%, and 2 for >94%.
	BS Prediction	2	This problem is relatively easy. You should be able to get a classification error at most 10% on the test data. Hence, I will allocate 0 point for anything more than 10% (>10%), 1 point for rates between 8% and 10%, and 2 points for rates below 8%(<8%).
Presentation	Video+vote	6	6 is reserved for the top five groups that survive two rounds of popular votes (see the next page for details); 5 excellent; 4 very good; 3 good; 1-2s below the bar.
	Total	13	

Note: Voting is meant to facilitate team building (and fun). Teams are encouraged to inspire each other in creating better strategies. Points will be subtracted for teams not participating in the voting. (see the next page for details).

Instructions for write-up

You will submit a report documenting your analysis. A typical format for the report would be:

a. One cover page and one Executive Summary page.
b. Review of the approaches that you tried or thought about trying.
c. Summary of the final approaches that you used to predict Value and BS, and why you chose those approaches.
d. Conclusions.

Among other things, points will be allocated for clear articulation of the approaches you considered, and the reason you chose the final approaches. **The main body of the report must be no longer than six pages (including the the cover page and the Executive Summary page).** However, you may also include a (up to) thirty page technical appendix with various computer outputs to justify the conclusions in your report. I will mainly look at the six-page main body, but might refer to the appendix if I see something unreasonable. The report should be named as **group_x_report.pdf**, for group x. All group members' names and student id numbers should be clearly indicated on the cover page. Also, please include the contact person's email (**and that person's email only**).

Business understandings

You should also consider: (1) in both prediction problems, would you argue if prediction accuracy or interpretation is more important? Why? (2) why do you think machine learning models might outperform Black-Scholes in terms of predicting option values? (3) can you argue from a business perspective that all four predictor variables should be included in your prediction (i.e., no variable selection is necessary)? (4) Are you comfortable about directly using your trained model to predict option values for Tesla stocks? Why?

Presentation and student voting scheme

Only the top ranked videos will be played in class. I will randomly divide all groups in five cohorts (I, II, III, IV and V). Each group will watch all group presentation videos of another cohort. For example, each group in cohort I will watch all videos in cohort II. The contact person of each group should email to the TAs (dso530.spring2024@gmail.com) their choices of top 3 (we do not distinguish among the top three) by **10 pm on Saturday April 20**. The email should be titled: "group x round 1 vote" (x indicates your group number). In this round, the voting criteria include clarity of the presentation, rigor of the approaches, and creativity. The TAs will aggregate the votes and announce the top three in each cohort. So there will be fifteen groups selected in the first round. The class will be informed the 15 finalists **late on Sunday April 21**.

On **Monday and Tuesday, April 22 and 23**, in each class, we will play videos of three selected groups, followed by my comments and Q&A. As we have five sessions, in total, I will play 15

group presentation videos on that day. Round 2 will consist of voting for the top 5 teams of the 15 finalists by the instructors, Prof. Xin Tong and Prof. Paromita Dubey, to determine which teams receive the full 6 / 6 presentation points. The remaining teams will get a maximum of 5 out of 6 presentation Points. In case there are ties, the TA votes will be the tie breakers. The results of round 2 votes will be announced by the instructors on **April 27**.

If your group does not participate in the voting, your group will be removed from other groups' votes (if there are any) and up to **1 point (0.5 for round 1, and 0.5 for round 2)** will be taken away from your presentation score. **No group should lobby for votes.**

Deliverables (no late submission accepted)

(1) Upload your **(up to) 10 min presentation** video in mp4 format (group_x_presentation.mp4). Deadline: **April 19 at 5 pm**. The contact person should submit your group video to: https://www.dropbox.com/request/Zrtso4z4YEe91dsUAko6 . Please double-check your video before uploading it. A video longer than 10 min faces penalty.

(2) Submit two files, i.e., your report (**group_x_report.pdf**) and prediction (**group_x_prediction.csv**). Deadline: **May 2 at 5 pm.** The contact person should submit to: https://www.dropbox.com/request/lkp15CTkWlsLyXORuXf1

Additional instruction for Dropbox submission: If you are not signed in with an email account on the browser, you will be asked to give your first name, last name and email address. If you already signed in with some email account, its information will be used automatically. **After the deadline, the link will be deactivated.** If you submit multiple times before the deadline, the last submission (with the same file name) will overwrite previous submissions.

Timeline:

1. April 19 at 5 pm, video upload
2. April 20 at 10 pm, round 1 vote
3. April 21, round 1 finalists will be announced
4. April 22 and April 23 in class, round 1 finalists' video play, Q&A and comments
5. April 27, round 2 results
6. May 2 at 5 pm, report and prediction

Group meeting with the instructor:

After April 26, the other groups (minus the 15 groups whose videos are commented in class) has an option to meet with the instructor for short consultations on specific aspects of their project. A schedule will be given later.