Data Analysis Project Peer Evaluation

STAT 420, Summer 2017, Xiaoming Ji (netid: xj9)

Table of Contents

Overview

STAT 420 is the first class I take for MCS-DS program. It's been a great and challenging experience working on this project and I personally learned a lot.

The subject we choose is "Residential Housing Price Prediction". It is a real problem with the real data set of **80** features and the prediction accuracy is quite demanding. I'm glad we finally figured some good models through hard working.

Postmortem

Besides statistical knowledge, the most I learned is how to better engage in such group project. This will definitely benefit my future study.

We are group of people in 4 different locations and timezones.

- Simon: ShangHai, China
- Tesa: HongKong/Sydney/San Francisco
- Martynas: Copenhagen, Denmark
- Shailender: Florida.

I feel the biggest challenge we have is to how to sync-up ideas and keep everyone engaged as much as we can. Skip the process, I'd like to list the **Postmortem** results I personaly would like to take if I'm doing similar data science group project again.

Members Should be Physically (or at least TimeZone) Closed

Today's technology can easily connect people anytime, anywhere and through any device. However, this doesn't means different location won't impact the efficiency of communication. Easy to allocate time that good for everyone and have a meeting/discussion is still very important. Such meetings are critical during the early time of the project to determine goals or when things go off the track.

Ideally, once a week face-to-face meeting is necessary for group members. Not only this is a way we can know each other better thus can cooperate better (I believe the physical socialize is still one of the best we can have in physical school), but also the most efficient way to make sure everyone is on same page.

Agree on Project Goal

We need to determine what goal we what to achieve in this project. Just to demonstrate the knowledge and skill we learned in the course? Or do diligence to find best model we can to solve the problem? The later is much more time consuming but we will also learn much.

Without such goal clearly agreed among members, we wont see people have same quality standard for the project.

For data science project, this also means we need to agree on **Test Approach**. For example: lowest LOOCV RMSE, AIC, BIC or most interpretable model.

Decide Report Skeleton and Clear Work Assignment

Once we decided the goal, we should first come out with structure of the report. Not just Overview/Methods/Results/Discussion. We need to have more detailed structure of the report such that each work can be assigned to each member unambiguously. For example: data pre-processing, outlier analysis, model selection/research, test, discussion etc. Try to make the work can be done in parallel to avoid bottleneck.

Besides study, online student usually has to spend more time on work and family duties. Without such clear assignments, its hard for every member to have good sense of ownership and fully committed.

Get Pre-processed Data Ready Early

Data is the foundation of any data science project. We need to have the pre-processed data ready as early as possible before we can explore the model and do evaluation. The mistake we made is we don't have unique pre-processed data set ready for everyone to explore until very late. This caused a lot of churns on inconsistent model selection and evaluation. Which caused lot of redundant work among group members.

Project Plan

Even for such short project, we still need a project plan to determine the schedule of each assignment. Assignment without due time is at risk of delaying the whole project. We should ask member to report progress on their progress and make contingent plan in case work is not done on time.

Give Feedback Early

We need to involve instructors early in case when we feel something is wrong, or can't make agreement among team or people is not engaged. We respect everyone's agenda and idea, but when things go out of control of the group, we should better let the instructor know ASAP to help us make appropriate adjustment.

Peer Evaluation

I'd like give thanks to everyone's work. Here is the review.

- Me/Xiaoming Ji (xj9)
  - Which parts of the project were worked on by that member
    - Almost every part. Specifically, nominated the project topic, completed project proposal and submission, come out with systematic data pre-processing approach, a systematic way to select predictors, find a way to optimize model for vif and loocv rmse, test against kaggle score, made discussions and appendix, and prepared the final report for team review, and report submission.
  - How well that member communicated with the team (**Score: 80**)
    - Apparently I'm the most active member in this group and emailed/slacked most. However, I still give myself relatively low score on this because as I mentioned in Postmortem, I feel I need to improve the way to communicate with the team and be more effectively , set right expectation and make everyone better engaged.
  - How well that member understood the course concepts (**Score: 100**)
    - I'm pretty confident on my understanding of this course. I also read other books and articles related to linear model (see appendix in the group project report)
  - Proportion of the project completed by that member (**Score: 77%**)
    - Yes, its a little unbalanced work and commitment among us. Completed: Data Pre-Processing, Model Selection, Evaluate with Kaggle, Discussion and Appendix.
- Martynas Sapoka (msapoka2)
  - Which parts of the project were worked on by that member

        - Discussed the project proposal, tried to find important features in the data set.
    - How well that member communicated with the team (**Score: 80**)
        - Not too often but also know to set low expectation.
    - How well that member understood the course concepts (**Score: 80**)
        - Seems able to perform some basic research on relevant topic.
    - Proportion of the project completed by that member (**Score: 3%**)
        - Demonstrated some effort but has no much contribution.
- Shailender Singh (ss49)
    - Which parts of the project were worked on by that member
        - Discussed the project proposal at very beginning.
    - How well that member communicated with the team (**Score: 60**)
        - No much communication with the team.
    - How well that member understood the course concepts (**Score: 60**)
        - Haven't demonstrated anything.
    - Proportion of the project completed by that member (**Score: 0**)
        - Really nothing.
- Tesa Ho (tnho2)
    - Which parts of the project were worked on by that member
        - Nominated project topic, worked on first version of the project proposal, presented early data analysis and model evaluation method.
    - How well that member communicated with the team (**Score: 90**)
        - Tesa is the second most active member in the group. Very active on trying to setup team meeting, proposing project topic and did what he said. We have some debate on the goal of this project, I inclined to spend time and find the best model while Tesa's major purpose is to finish the work. But that's fine and I respect people's true intention of taking this class.

    - How well that member understood the course concepts (**Score: 90**)
        - I believe Tesa has good understanding of this course. He is proficient on using R to visualize data and do model analysis. Although, I do feel he could spend more time to think the problem systematically and come out with a thoughtful solution. Data science project is not always an easy task, it needs patience and endurance to make the best results.
    - Proportion of the project completed by that member (**Score: 20%**)
        - Completed introduction, correlation scatterplots and histograms for data analysis.