**Case study**

Over the semester, we have examined different modeling technologies. As we are approaching the final presentation, one of the challenges we are facing is how to combine different modeling technologies and work as a group. In this assignment, you are asked to read an article written by a team that won the 1st place in a Kaggle competition for predicting credit default. Please read through the article and answer the questions below. Hopefully, you will learn something beyond the predictive models.

*Link to this article:* [*Ranking #1 on Kaggle for Predicting Consumer Debt Default*](https://blog.nycdatascience.com/student-works/kaggle-predict-consumer-credit-default/)

*Link to the competition:* [*Give Me Some Credit*](https://www.kaggle.com/c/GiveMeSomeCredit)

1. What is the purpose of building a credit default model? Why is such a model usually very complex?

To build a model that borrowers can use to help make the best financial decisions, and for the lenders to foresee when a borrower might lead into financial distress. The intent is to improve on the state of the art in credit scoring by predicting probability of credit default in the next two years.

A banks and various lending institutions are constantly looking and fine tuning the best credit scoring algorithms out there. It is the model banks use to determine whether or not a loan should be granted. With this information, banks can make better decisions and borrowers can also do better financial planning to mitigate possible default status in the future.

2. What is the definition of ensemble models? What ensemble models have been introduced in our class?

Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model.

Regression, Decision Tree, Random Forest have been introduced in our class.

3. Can you do a SWOT analysis for *your* final project? List the strengths, weaknesses, opportunities, and threats you are confronted with.

Strengths - Leveraging what we already know

A synergistic team with very complementary skills and experiences

Lessons learned from a previous Kaggle challenge

Impact of drinking on students。

Weaknesses - Areas that we need to improve upon

There are a lot of variables in the data, and it takes me a lot of time to sort it out.

Learning while executing will slow down the entire process

Sparse resources on some of the newer technologies employed

Opportunities - Chance to apply & learn from experience

Learn how to strategize, optimize, and fine-tune models, algorithms.

Parameters to achieve the best default prediction possible

I will learn more modeling methods and analysis skills from it.

Threats - Risks that we need to mitigate and manage

Small dataset size presents tremendous challenges on generalization that would impact modeling and ultimately, prediction accuracy

I have never done a similar survey. It's going to be a huge challenge for me.

4. What approaches have been taken by the winning team in the stage of exploratory data analysis before running a predictive model?

In the stage of exploratory data analysis, they tried different imputation methods, including KNN, mean, random and median, and PCA. But all of these may not work well for them. Then, as almost every column except ‘age’ in this data set has some problems with the outliers, they detect the 95% percentile-based outlier, median-based outlier and standard deviation outlier and do a majority vote for each column. If two out of the three methods agree that a certain point is an outlier, they identify this point to be an outlier. Then they replace the outlier either by the median of the column or the least outlier depending on the column.

5. What models are included in the blend of Voting and Stacking models that the team used to secure the 1st place?

Two Gradient Boosters, two Random Forests, one AdaBoost, Restricted Boltzmann Machine (RBM) and the Neural Network algorithm.

6. What is the definition of a “tuning parameter”? Please list the tuning parameters introduced by our class.

Almost every Machine Learning algorithm include some number of hyper parameters, also called tuning parameters.

These hyper parameters are different from regular parameters in that they are not part of a model, therefore not tuned automatically through model fitting; they are tuned using a separate process altogether. Some examples of hyper parameters include the regularization parameter lambda for ridge and lasso regression, the C term for Support Vector Machine, the number of trees for tree-based methods.

7. What are the most important lessons you learned from this article?

This article helped me learn some valuable lessons about machine learning, predictive modeling, and teamwork. I'd like to explain SWOT analysis, which allows teams to better organize ideas and immediately put team members on the right path. SWOT analytics helps teams take advantage of their greatest strengths, understand their weaknesses, take advantage of opportunities, and be aware of any threats to their business. It is a good choice for a team to conduct a self-evaluation before the start of the project.