Sentiment and Grade Prediction from RateMyProfessor Reviews

MARK MALYSA, mbm206, section 07

May 2025

1 Introduction

Toward the end of every semester, every student is tasked with picking their class schedule for the following semester, and with this choice comes the choice of which professor they prefer. Besides testimonies from past students, many students flock to RateMyProfessor.com, an anonymous website where students can leave reviews about professors. In addition to actual comments left by students, the student can enter a rating, difficulty, For Credit, Would Take Again, grade, Textbook, Attendance, and Online Class. The problem with these reviews is that they can vary drastically, making it difficult for students to predict which professor is better. Additionally, many reviews are conducted during the semester, so a final grade is not available.

What this project will aim to do it to conduct a sentiment analysis on the review comments left on the professor, then it will use other features such as difficulty and rating to predict a grade for the class. This way, students are able to have a tool to accurately provide a predicted grade for any given professor and any given review.

2 Motivation

The project is important to me because I am currently a student who has gone through the stress of trying to pick the best professors. Many times, I feel as if my choice was not made accurately or that the few reviews I read were not good testimonies relating to the teacher. I want to see if it is possible to train a model using data gathered from Rate My Professor and see if it can be used to accurately predict grades.

Some prior questions that can be presented is just how important is the sentiment of a review? Although this project will focus on an educational view, this style of work can be applied to any review system. It will certainly be interesting to see the importance of each feature and how they correlate with the grade predictions.

Prior works related to this topic have been done, as sentiment analysis of reviews is not a new idea. There are many prices of workout programs out there that have classified reviews as positive or negative, and even some that have been done with professor-based reviews. However, I was not able to discover much about predicting grades using these sentiment analyses as features.

3 Method

The dataset that was used was "Rate My Professor Reviews 5C Colleges" from Kaggle (https://www.kaggle.com/datasets/tilorc/rate-my-professor-reviews-5c-colleges). This is a public data set that consists of web data that was scraped in Spring 2022, and each review consisted of eight variables: course ID, Quality, Difficulty, For Credit, Would Take Again, Grade, Textbook, Comment, Professor, and Department. The actual file structure was a JSON file that consisted of lists of reviews, where every list corresponded to a specific professor. Each of these reviews contains text data, numerical data, and categorical data. The features that are particularly important to this project include the comment, the rating, the difficulty, and the grade; however, all were considered.

In this project, two different models were used. The reason for this was poorer performance than expected from the first. Hence, another model was brought in and tested to arrive at better results. The first was a Random Forest model, which was chosen due to some prior research that involved viewing other projects pairing similar features to predict an outcome. A random forest builds many different decision trees, allowing all of them to perform, and then averages the results. This has its advantages as it handles non-linear relationships, and it allowed me to extract feature importance. This second advantage was an important quality, as it was important to see which chosen feature was most important to look at when trying to predict grade. The second model that was used was a gradient-boosting regressor, primarily due to its often having a higher accuracy than random forests, since it builds trees sequentially and learns.

3.1 Problem/Feature Space

The problem space and task was defined as predicting the grade for each individual review on a grade point average scale. First we took the target variable, being grade, and converted this from a letter grade to a gpa scale. Concerning the feature space, each review can be represented by a feature vector that contains the sentiment data, the numerical data, and the categorical data. The numerical data consisted of the difficulty rating, the overall rating, review count, and the average rating and difficulty for the professor. The categorical data consisted of the department, which was also converted to numerical values.

3.2 Implementation/Evaluation

To implement this, first pandas were used to not only load the dataset, but they were flattened so that each row only represented one review. Next, the data was cleaned by handling the missing values by either filling them with averages or dropping the rows completely. In addition, duplicate rows were removed and all the text was cleaned and standardized. The sentiment scores were then calculated using a sentiment model from the NLTK Python library. Following this, to finish processing the data, average scores were calculated for each professor, and letters were converted to a 4.0 GPA scale.

To follow this, the data was split into two separate groups, the training data and the test data. The training data was then used to train the two models that we selected, and these models were then evaluated on the test data. The models were evaluated using metrics that included Accuracy, F1 Score, Root Mean Squared Error, R^2, and then the feature importance scores were extracted from the model. To finish this multiple visuals were created to show feature importance, correlation, actual vs predicted, and other relevant information.

4 Results

4.1 Sentiment Analysis

As seen below in the figure the sentiment analysis showed primarily positive results giving many of the reviews sentiment scores very close to one. The graph displays a left skew showing a high median for the overall sentiment scores that the model predicted.

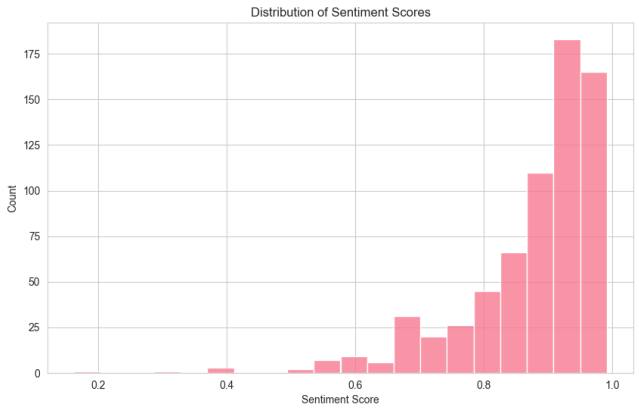

Figure 1: Test Sentiment Distribution

The model achieved a training accuracy of 0.7370 and a F1 score of 0.8446,

while also getting a test accuracy of 0.8000 and a F1 Score: 0.8880. Looking at the classification report it can be seen that the model had a very high recall rate for positive reviews being 1.00 but a very low recall rate for negative reviews. This means that although the model successfully identified all the positive reviews it struggled with the negative reviews which can be due to the class imbalance in the data. The data that was provided heavily favored positive data which can explain these results. This can further be seen with a false positive rate of 0.20 and a false negative rate of 0.0. Below one can see the correlation between the sentiment scores and the grades for the test data.

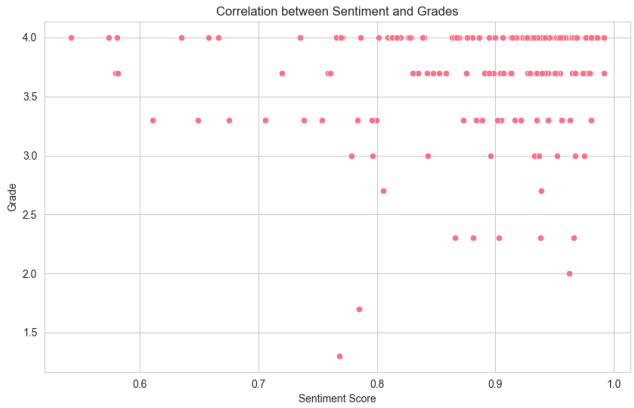

Figure 2: Test Grade Correlation

4.2 Random Forest Model

The Random Forest model had a low training RMSE of 0.098 which differed drastically from the test RMSE of 0.259. This showed that the model fit the training data very well but may have overfit the test data. The mean absolute error (MSE) was 0.144 which says that the model was on average 0.144 grade points off from the actual grade. However critically the model did not perform better than simply predicting the average grade for all students, shown by the negative R^2 value of -0.085.

4.3 Gradient Boosting

Comparing the Gradient Boosting model to the Random Forest model, it had a higher training RMSE of 0.208 but a slightly lower, therefore better, test RMSE of 0.254. This difference is not significant but it does suggest the gradient boosting model was a better fit. The test MSE was 0.127, meaning that the model's predictions were on average 0.127 grade points off from the actual grade. This is better than the Random Forest Model showing more close to accurate

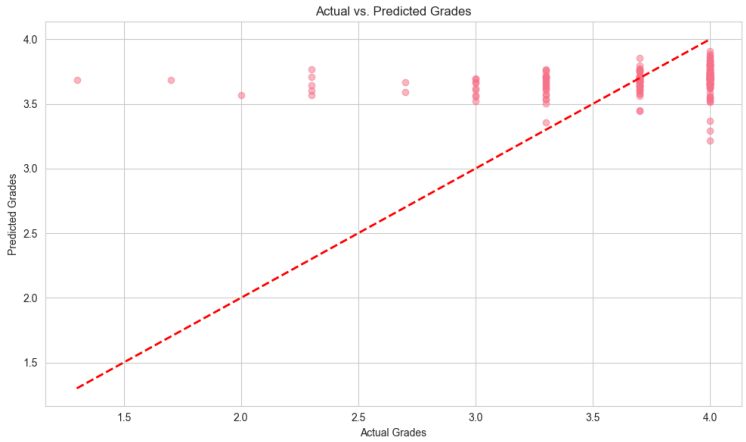

Figure 3: Random Forest Actual Vs Predicted

results. Al thought these results were optimistically better than the that of Random Forest the model still produced a negative R^2 value of -0.046. This indicates that this model also did not perform better than simply predicting the average grade for all students.

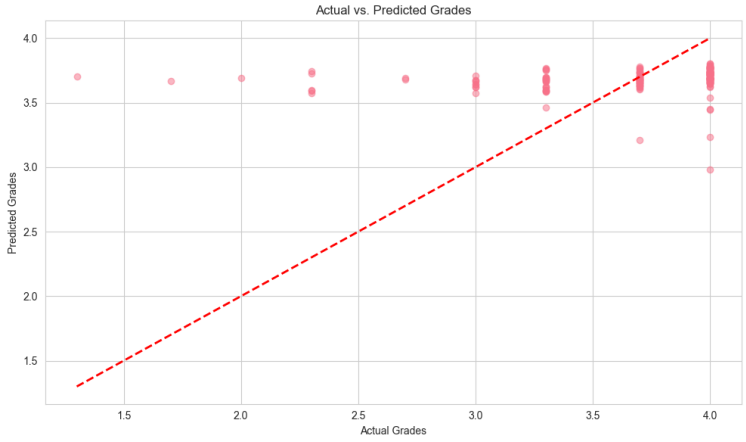

Figure 4: Gradient Boost Actual Vs Predicted

4.4 Feature Importance

Both models presented a results with the same top feature of Sentiment Score being the most important. The Random Forest Model had an important of 0.34 and the Gradient Boosting Model had an importance of 0.26 for Sentiment

Score. For the Random Forrest Model the next strongest predictors were shown to be average rating, average difficulty, review count, rating, and difficulty. As seen from the graph below Sentiment Score is shown to have a much higher importance compared to the rest of the values. The gradient boosting model gave the next strongest predictors of rating, average rating, average difficulty, difficulty, and review count. This order differs from the other model however, as seen below, sentiment score does have the highest importance but it doesn't differ as significantly as with the other model.

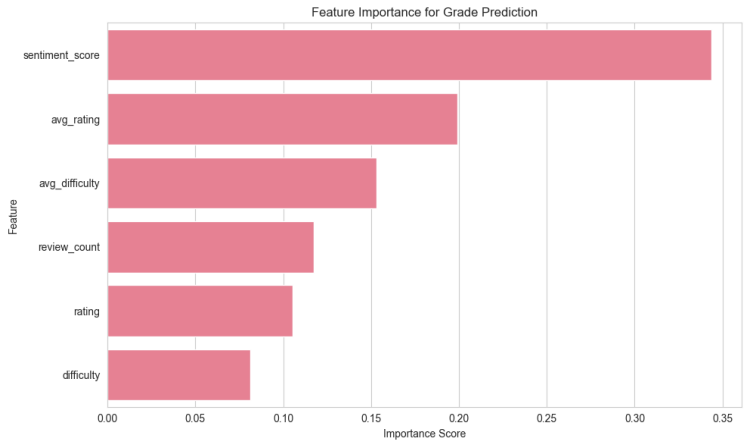

Figure 5: Random Forest Feature Importance

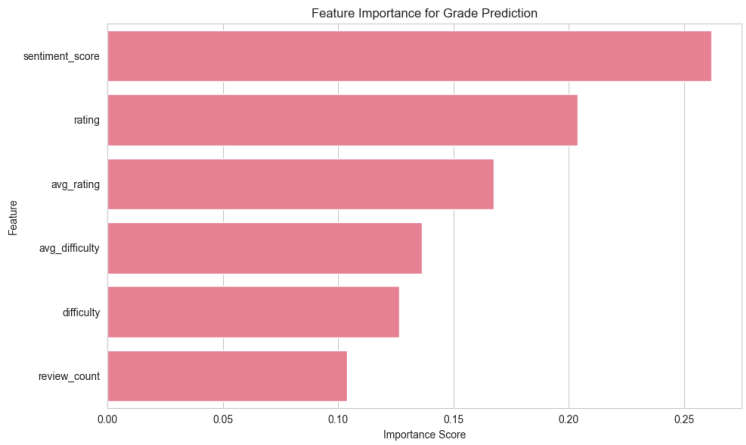

Figure 6: Gradient Boost Feature Importance

4.5 More Visualizations

As seen below the correlation matrix shows that the highest correlation is seen between the features rating and sentiment score.

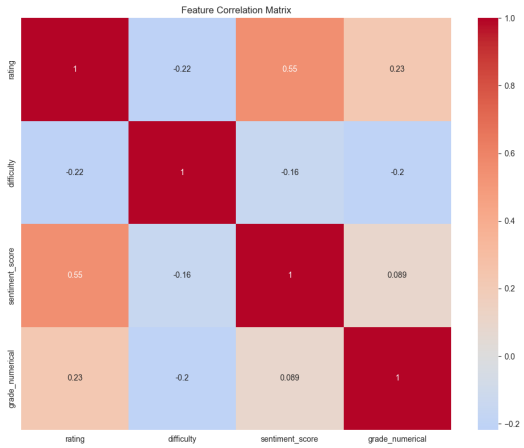

Figure 7: Random Forest Test Correlation Matrix

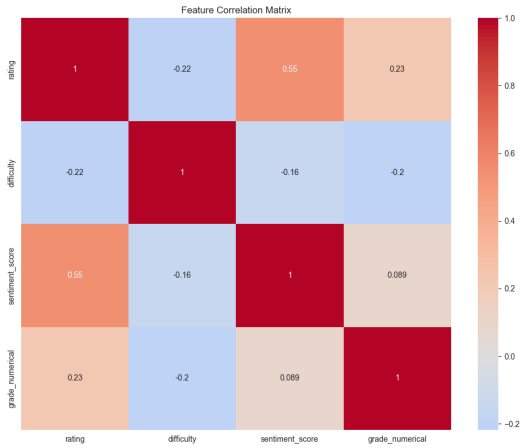

Figure 8: Gradient Boost Test Correlation Matrix

5 Discussion

Overall, the model did not perform as I expected it to and I did not expect to see the outcome that prospered. The sentiment analysis model performed perfectly

when predicting positive reviews but due to a class imbalance did not perform as well with negative reviews. For the two models I originally settled on using a Random Forest model for its advantages but the model did not perform as well as expected. The model suggested overfitting shown by the lowered training error compared to the test error. This resulted in shifting to use a gradient boosting model which did outperform the random forest model only slightly. What this may suggest is that possibly a different model should be attempted or possibly the data set was not the best to train these models. In the future this research can be used to better consider plans to extend the thought of sentiment analysis on grade prediction.

My final result did differ from what I originally planned and what I originally sought out to display. I first wanted to scrap the rate my professor reviews however this proved to be very difficult and time consuming due to the way the website is constructed. Dealing with the features and correlations themselves I first thought that the sentiment analysis score would have the largest correlation with grade prediction. However this was far from true. Using feature importance I was able to confirm my original insight that the sentiment analysis would have the largest importance.

The project sought out to try to predict grades for professors based upon rate my professor reviews to better help students when picking course. The insights showed that using a Random Forest and Gradient Boosting model to predict so may not be the best case and the sentiment may not be correlated with the grade at all. Instead the sentiment is higher correlated with the rating given to the professor. In the future I would like for this project to be continued and further explored. It would be interesting to see this idea presented using different models or used on different sectors besides education. This would possibly show or reveal just how influential sentiment is to the actual grading process of an establishment or individual.