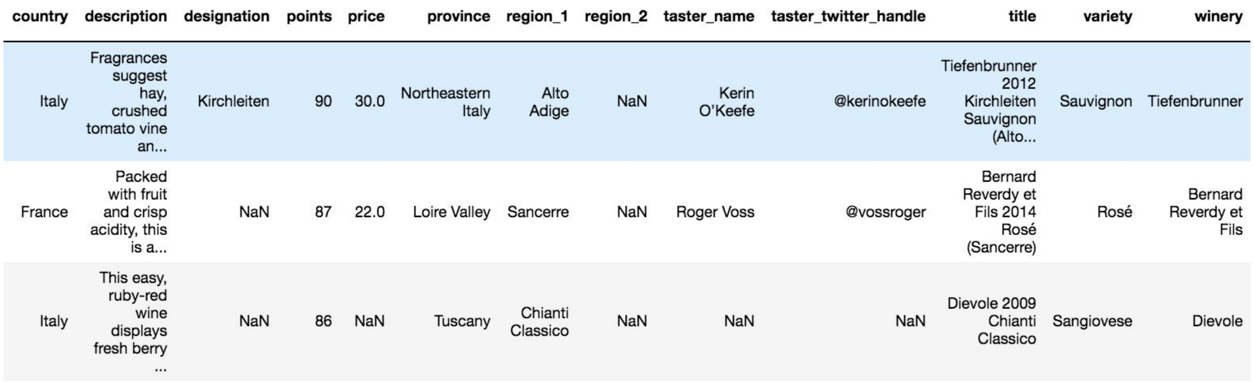
Introduction

In this report, we explored how predictive 13 fields of data can be towards a wine rating online shop. Each and every field tells us something about either the wine type or a wine taster.
Our goal is to choose the most predictive field and build a good model as a baseline.



	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	title	variety	winery
	Italy	Fragrances suggest hay, crushed tomato vine an...	Kirchleiten	90	30.0	Northeastern Italy	Alto Adige	NaN	Kerin O'Keefe	@kerinokeefe	Tiefenbrunner 2012 Kirchleiten Sauvignon (Alto...	Sauvignon	Tiefenbrunner
	France	Packed with fruit and crisp acidity, this is a...	NaN	87	22.0	Loire Valley	Sancerre	NaN	Roger Voss	@vossroger	Bernard Reverdy et Fils 2014 Rosé (Sancerre)	Rosé	Bernard Reverdy et Fils
	Italy	This easy, ruby-red wine displays fresh berry ...	NaN	86	NaN	Tuscany	Chianti Classico	NaN	NaN	NaN	Dievole 2009 Chianti Classico	Sangiovese	Dievole

Methodology

Most columns are categorical.They are either filled with a large percentage of null values or a high number of unique values. One suitable path to explore will be to one-hot encode all categorical columns. However, with columns like winery having 5460 unique values, we will end up with a 5460 columns for just that feature and hence introducing the curse of dimensionality.

For these reasons, we will focus on using the description and points field to build our model. I feature engineering some extra fields for my analysis. These include features such as the Length of text, number of words, number of non stop words and average word length. In summary, given the description of a particular type of wine, the above features are generated on the fly and added as columns for analysis. This tends to work out great!

One very important part of our analysis is the problem formulation. Technically, our problem is currently formulated as a regression analysis problem. Thinking about it carefully, we realize that predicting the exact point is not what we are interested in. Instead, we want to be able to categorize the wine types into groups that make business sense. We want For this reason, converted the points column into 4 categories:

1. Good - (80 - 84)
2. Very Good - (85 - 89)
3. Outstanding - (90 - 94)
4. Classic - (95 - 100)

We have now successfully reformulated the problem into a natural language classification problem. We can now go ahead and apply the general steps in data processing of NLP problems to our problem. These include,

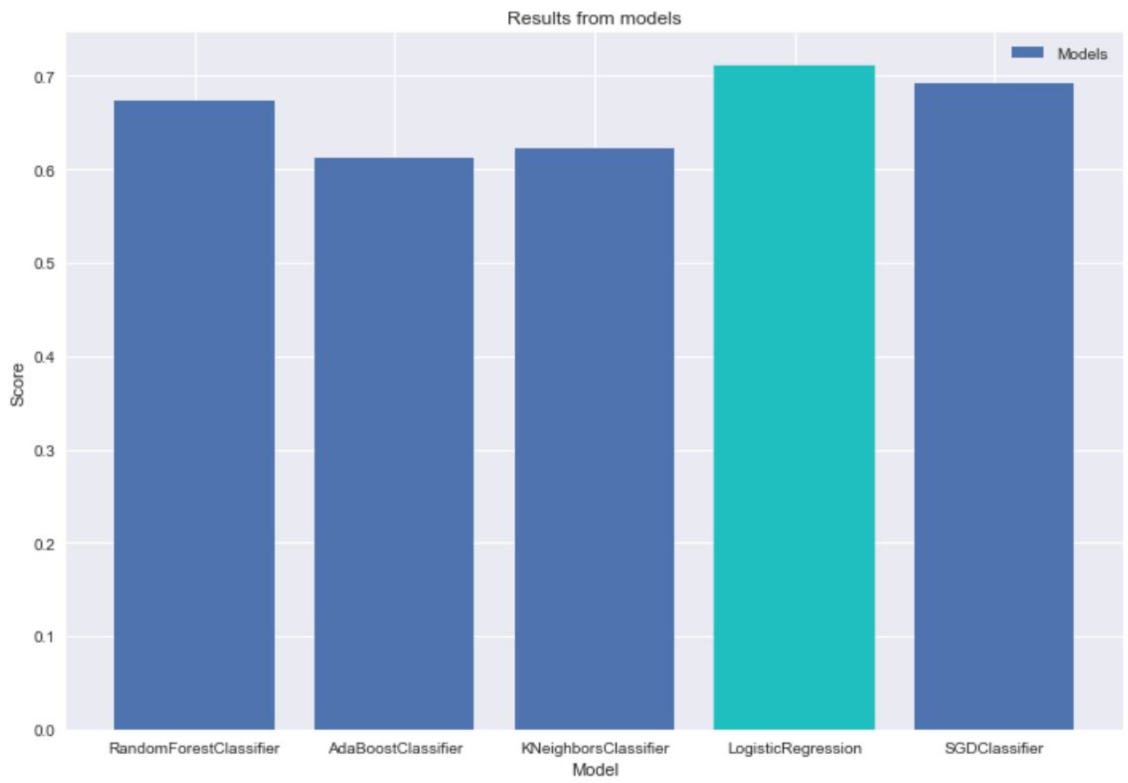
1. Converting description field to lowercase
2. Removing punctuations
3. Expanding contracted forms
4. Removing stop words
5. Tokenization
6. Vectorizing the tokens.

However, we also have some engineered numerical columns we must take care of.

To make the code modular and expressive, I resorted to scikit-learn pipelines for my analysis. I put most of the processing steps above into the pipeline and built models on five different algorithms. Using the accuracy score on the test set, I chose the most predictive algorithm. I further enhanced it by doing a grid search on a 5-fold cross validation. It resulted in a slight increase in the accuracy.

Results

After evaluating the pipelines on 5 different algorithms, logistic regression turned out to be the most predictive. It gave an accuracy of 71%. The diagram below is an illustration of the performance of the various algorithms.

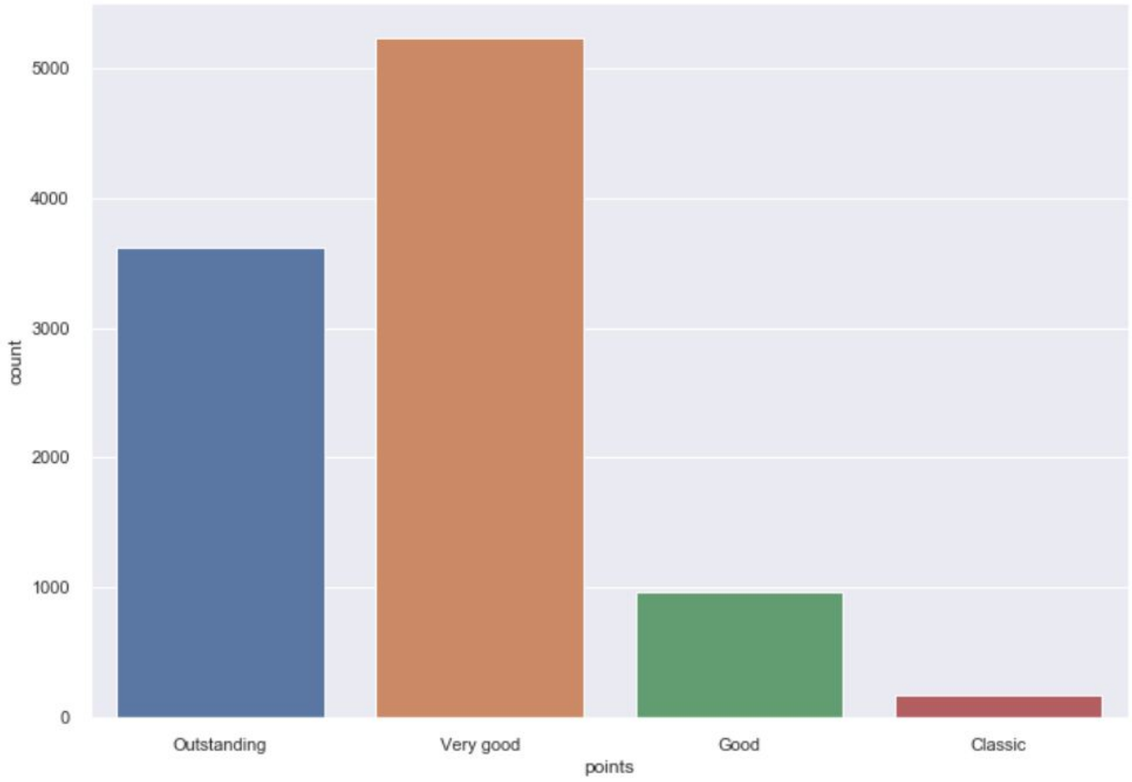


Interpretation

Drawing conclusions from just accuracy would be rather naive. Let’s expand to see how predictive logistic regression was on the various categories in the dependent column. Here, we’ll use the classification report.

	Classic	Good	Outstanding	Very Good	Weighted avg
f1-score	0.188679	0.187500	0.723243	0.733209	0.665181
precision	0.384615	0.750000	0.675740	0.705566	0.692555
recall	0.125000	0.107143	0.777929	0.763107	0.691500

It can be observed from the table above that, the logistic regression algorithm is relatively accurate in predicting the ‘Very Good’ and the outstanding class. This stems from the data imbalance that resulted after we converted our points column to 4 categories. It was very poor at predicting the Classic category. This is explained in the relatively low sample size of wines belonging to that category. Here’s a diagram of the class imbalance.



Clearly, it can be seen that, “Outstanding” and “Very good” form the largest group in our dependent column while “Good” and “Classic” were really minimal.

Suggestions for Improvement

The current model is just a baseline. I see a lot of ways to improve the performance of our models. To improve the current model, I suggest the following:

1. Gather more data
2. Tackle class imbalance using SMOTE
3. Build an ensemble
4. Do a randomized search on the current logistic regression algorithm
5. Extract and use year of production form the title of column as a feature
6. Combine description and title columns and go through the same steps highlighted above
7. Use deep learning