Knight FinTech Internship

Model proposal for “The Wine Land” task

MAY 2020

ISSUED BY

R. Senthil Kumar

iam.rsk@outlook.com | 7339659559

Background

“The Wine Land” is an online store that specializes in selling different varieties of wines. As the company’s data scientist, I have been assigned the duty to Leverage the “reviews” data and draw actionable insights from it. A predictive model for predicting the wine “variety” has to be built using Machine Learning techniques

The objective of this Proposal is to showcase an CNN system that will provide the best overall value to “The Wine Land” . While price is a significant factor, other criteria will form the basis of our award-decision (as more fully described in the Feature Extraction section of this Request for Proposal below).

Evaluation Factors

The best model has been selected by taking into consideration the following factors

1. Accuracy of the model
2. Precision - Recall measures
3. The columns taken into consideration
4. Ability to deal with local text-patterns

Based on the above factors, models of different configurations were built and trained, but only 2 models managed to provide satisfactory results:- ANN-dense and CNN-dense model. The CNN-dense is chosen as the best suited model.

The model

The dataset provided shows us that we will need to implement supervised learning. Out of several approaches available to us (such as decision trees, logistic regression, ANN), the Convolutional Neural Network gave us the best results. Following are the features of the proposed model:

1. 2-layered Convolutional Neural Network (as base) with Embedding layer (at bottom) and MaxPooling layers (in between adjacent CNN layers), and a Dense layer (on top)
2. RMSprop optimizer and categorical_crossentropy loss function
3. Architecture: Input layer->CNN base->28-node Output layer

```
Model: "sequential_19"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_15 (Embedding)     (None, 20, 16)            280176
_________________________________________________________________
conv1d_42 (Conv1D)           (None, 18, 16)            784
_________________________________________________________________
max_pooling1d_20 (MaxPooling (None, 6, 16)             0
_________________________________________________________________
conv1d_43 (Conv1D)           (None, 4, 32)             1568
_________________________________________________________________
global_max_pooling1d_11 (Glo (None, 32)                0
_________________________________________________________________
dense_24 (Dense)             (None, 28)                924
=================================================================
Total params: 283,452
Trainable params: 283,452
Non-trainable params: 0
_________________________________________________________________
```

Figure 1: Model Summary

Feature Extraction

The dataset is a rich source of data: 80000+ rows, each featuring 12 attribute experiences. In the first look, it might look that the price attribute can help us in logistic regression, but the non-linearity of variety rules that feature out. Features like country, province, region_1 and region_2, winery might also look helpful, but the fact that the same vinery may produce several varieties of wine rules those features out as well. Review_description is a rich source of information, but the varying lengths and the amount of unique words used throughout the reviews makes it complex for any model to take a firm stand.

The review_title is hence chosen as the primary feature for feature extraction, due to the following reasons:

1. Review titles sometimes contain a direct/indirect indication of the variety of wine being reviewed.
2. Less unique words- we only need to deal with 17510 top frequent words, whereas in review_description contains 50000+ unique words!
3. The max length of review_title is 135 and max words used in review_title is 19, which is handle-able indeed!

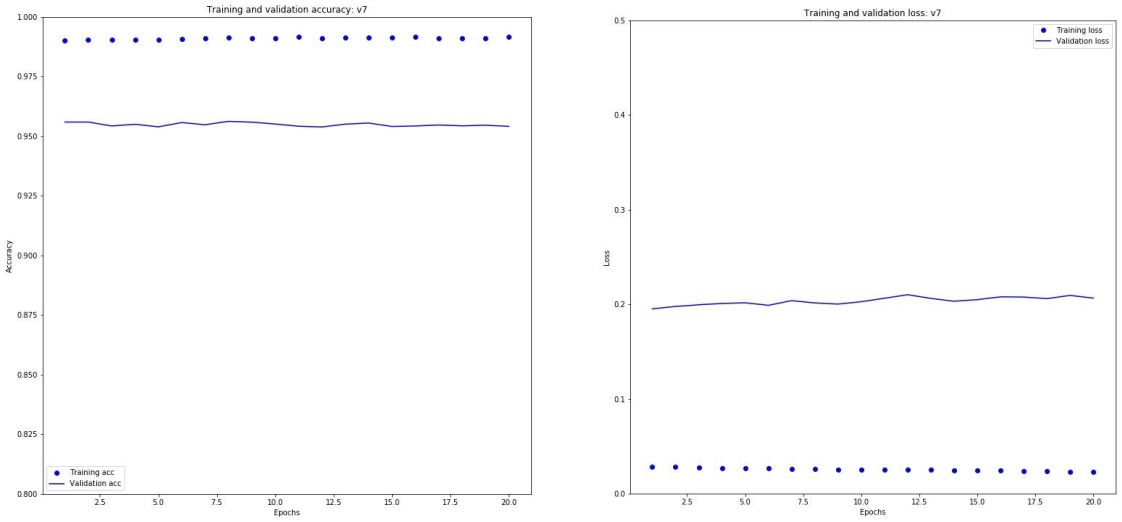
How is this feature EXTRACTED?

STEP 1: Create a dictionary of all unique words used in review_title, and assign ranks based on frequency of occurrence.
STEP 2: Convert all the words in each review into an array of integers, where each integer is assigned with respect to the word's order value in the dictionary created.
STEP 3: Join arrays of all reviews and make it as a list. Then, convert the list into a 2D tensor by padding each review-array to the max-word-length of review_title (20 in this case).

Outcomes

Upon successful training of the model stated above, we have achieved the following metric-values as a result:

- **Accuracy (on train):** 99.15%
- **Precision (on train):** 99.98%
- **Loss (on train):** 0.0231
- **Average time-per-step:** 51 us/step



Accuracy and Loss learning-curves for the trained CNN model

Data Visualisation Graphs

Following are the data visualisation graphs that contribute to creating several useful actionable insights:

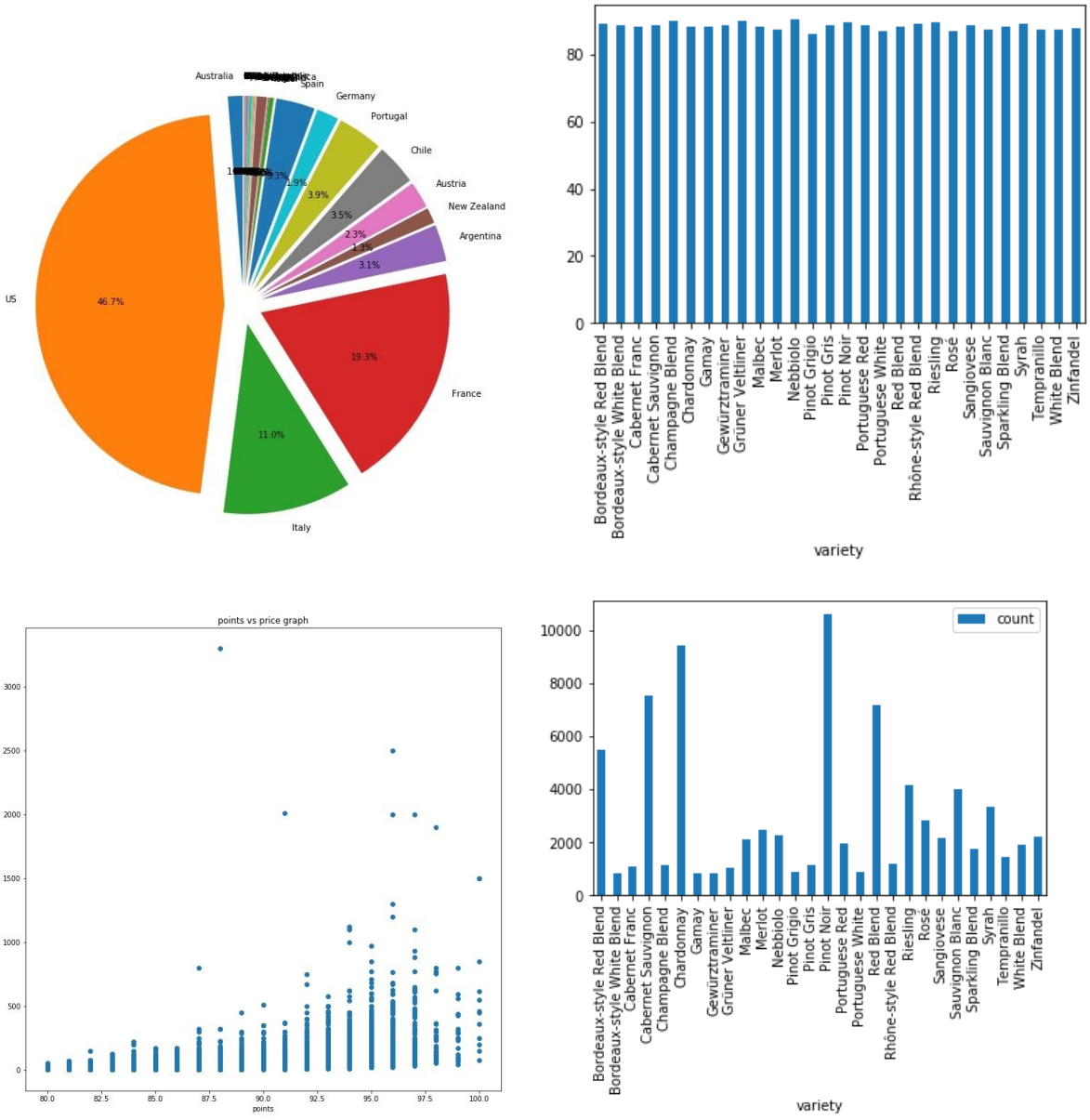
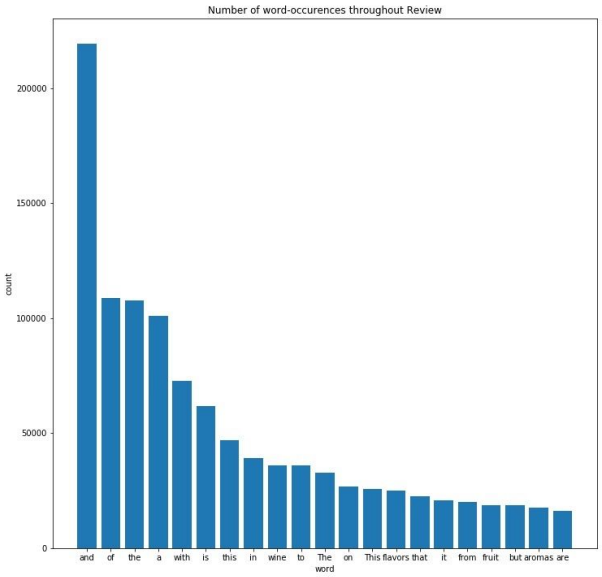


Figure 2: Visualisation Graphs : Part 1

[Left to Right, Top to Bottom]: 1. Percentage-wise share of each country on wine purchase, 2: Average points secured by each Wine variety, 3. Points vs Price graph, 4. Count of each variety in the review database

Testarossa	175
Louis Latour	168
Williams Selyem	165
Chateau Ste. Michelle	163
Georges Duboeuf	163
Wines & Winemakers	142
DFJ Vinhos	131
Concha y Toro	112
Columbia Crest	112
Kendall-Jackson	100
Siduri	99
Gary Farrell	98
Lynmar	98
Albert Bichot	94
Jean-Luc and Paul Aegerter	92
Montes	90
Chanson Père et Fils	89
Fess Parker	85
Henri de Villamont	85
Martin Ray	85



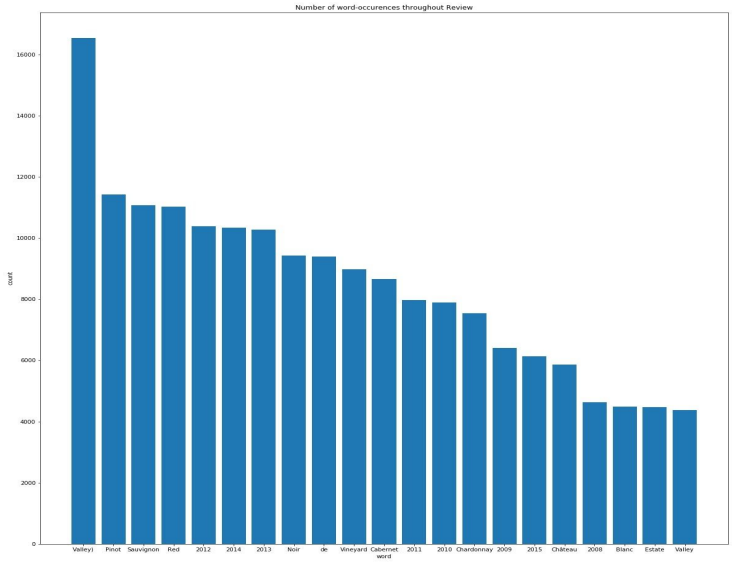


Figure 3: Visualisation Graphs : Part 2
[Left to Right, Top to Bottom]: 5. Top 20 prominent Wineries, 6 Occurence of top 20 words in the review_description, 7. Occurence of top 20 words in the review_title

Actionable Insights

The top Actionable Insights that are derived after taking into consideration the Data Visualisation Graphs and models, are enlisted below:

1. Increase Supply on Countries that Demand more

People from the USA, Italy and France alone contribute to more than 75% of the records in the review dataset. Hence it would be a good idea to establish local branches/ regional headquarters in these countries. Establish links with local wineries and maximise business by meeting demand-supply needs in these countries. For countries that take less share, such as India and Australia, focus on targeted marketing and advertising.

2. Amplify sales on high-demand wine category

From the Data Visualisation Graphs, it is clear that Red Blend, Port Noir and Charondarray are among the top selling wine varieties. Hence, necessary steps must be taken to speed up production of these varieties.

3. Supplier retainment - getting closer to prominent wineries

The success of a business solely relies on the quality of the product it sells. In this case, choosing promising and prominent wineries as key suppliers for the company is a crucial step. From the Graphs, it is clear that Testarossa, Chateau Ste. Michelle and a few others are among the most prominent wineries (by count). Hence, these wineries must be retained to maximise supply, and ultimately, profit.

4. Increase quality for wine varieties that have less-mean-point value and maximise quantity for wine varieties that have less-mean-point value

Mean point score is obtained for each variety. From the score graphs, we can clearly predict the taste-score for each variety! Hence, for each variety that has less score, quality improvement must be undertaken. For varieties that have a high score, supply must be maximised.

5. Fine tuning price and points

The higher the satisfaction point, the higher should be the price. Greater point scoring varieties will have greater demand and hence it will be a revolutionary decision to alter the price with respect to points in a close-to-linear manner. In order to achieve greater revenue in a relatively shorter timespan.

6. Process the most used words to get a fine idea on each variety

Words like 'a', 'and', 'the' are the most used words in review_description. On the other hand, Words like 'valley', 'red', '2007' are the most used words in review_title. A study on the usage of these words in review_title and review_description will help data scientists working in the firm to generate better and more useful results. Although this insight may not directly contribute to business expansion, it focusses on the company's R&D utilization and thus has an indirect impact.

The Alternates

Apart from the 2-layered CNN model, the conventional ANN-dense model also performs exceptionally well:- accuracy of 95%. But the model is not chosen to be the best-fit-model due to its inability to understand the close-word patterns. CNN networks have the ability to detect, learn and remember patterns that occur locally. But ANN-dense model can only predict an output for a given input combination, thereby missing the pattern factor. Images and Text data usually have recognisable pattern features, which the conventional ANN model fails to learn.
Nevertheless, this model can be considered as a base for future works to be done upon.

Inference

A very precise model has been built, with accuracy close to 98 percent. This model can be expanded to industrial fields and implemented in real world scenarios to attain greater heights.

"... for a brighter tomorrow!"