



# **Knight FinTech Internship**

**Model proposal for “The Wine Land” task**

MAY 2020

**ISSUED BY**

**R. Senthil Kumar**

[iam.rsk@outlook.com](mailto: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 ANN 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 Evaluation Factors 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 overfitting

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 model. The ANN-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), the Artificial Neural Network gave us the best results. Following are the features of the proposed model:

1. 2-layered Dense Artificial Neural Network
2. RMSprop optimizer and categorical\_crossentropy loss function
3. Architecture: input layer->32-node hidden layer->28-node output layer

```
Model: "sequential_2"
```

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 32)	560352
dense_2 (Dense)	(None, 28)	924

```
Total params: 561,276  
Trainable params: 561,276  
Non-trainable params: 0  
None
```

Model Summary

ANN model, input->hidden(32)->output(28) layers

The reason behind the remarkable performance of this model is because the model's structure is not complex and hence it easily adapts to the learning patterns of the dataset.



## 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 winery 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.

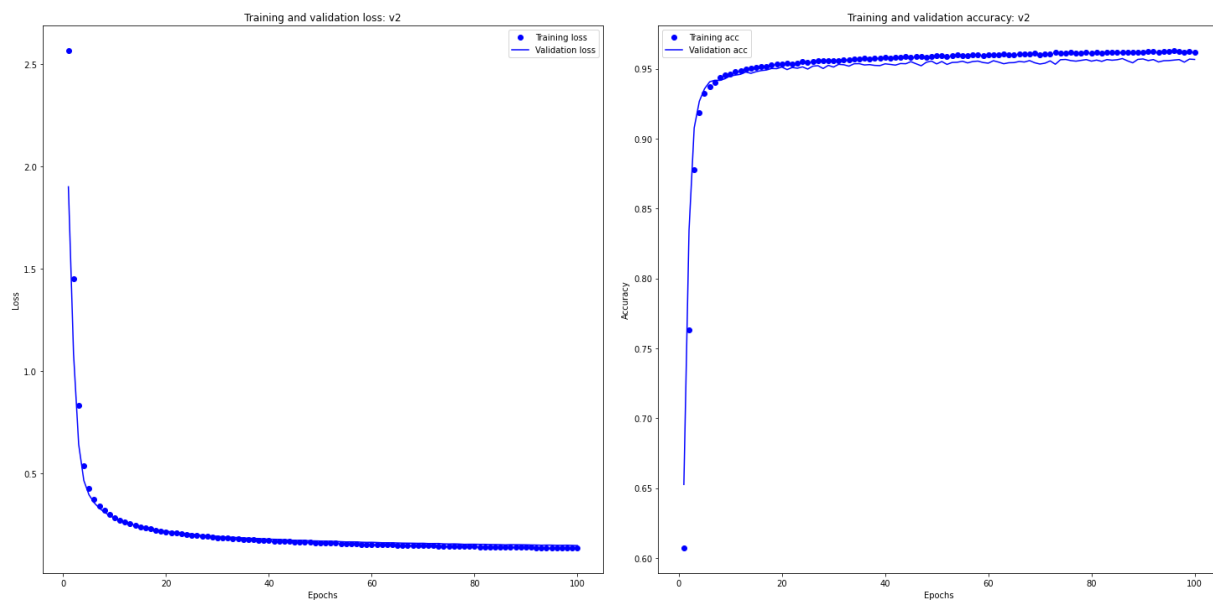
**STEP 2:** Only select those words that occur more than once to reduce complexity

**STEP 3:** convert each review string into a numpy tensor that says if a particular word x is present (1) or absent (0).

# Outcomes

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

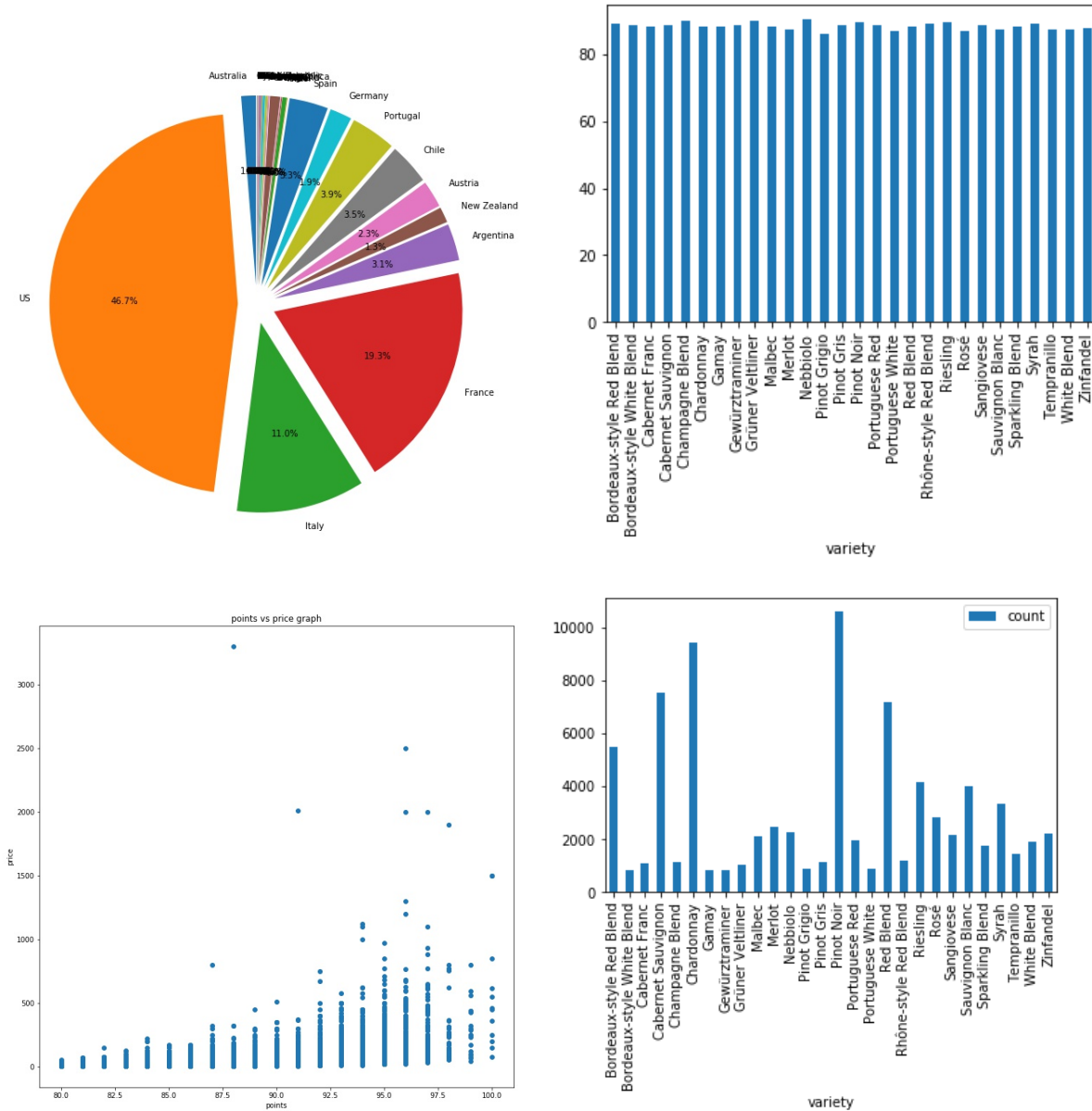
- **Net accuracy:** 98.96%
- **Precision:** 96.81%
- **Recall:** 95.60%
- **Loss:** 0.1477
- **Average time-per-step:** 120 us/step



Accuracy and Loss learning-curves for the trained model

# Data Visualisation Graphs

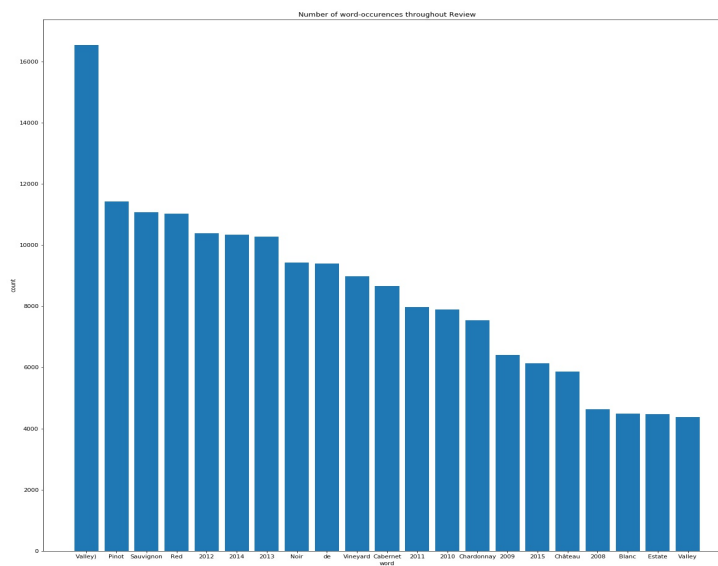
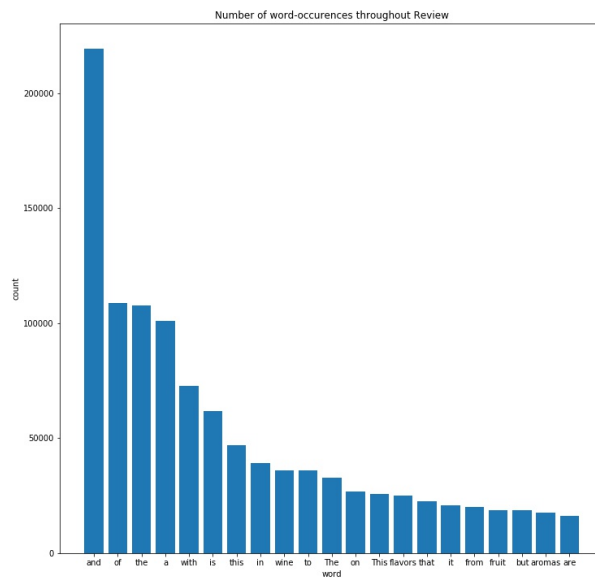
Following are the data visualisation graphs that contribute to creating several useful actionable insights:



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



## 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 conventional ANN-dense model, the 2-layered CNN model also performs exceptionally well:- accuracy of 90%. But the model is not chosen to be the best-fit-model due to its recall factor. This recall of CNN model fluctuates rapidly, with values moving close to zero after 20 epochs.

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!”*