A Comparison of Naïve Bayes and Random Forest on Airline Customer Satisfaction Prediction

Yuichi Kuriyama

Brief description and motivation of the problem

Due to the fierce competition in the saturated market in the aviation industry, Airline companies place more emphasis on the customer retention. In order to maximise the retention rate, the optimization of customer satisfaction plays a central role in further success in the foreseeable future. The aim of this study is to create the customer satisfaction prediction model of American aviation company from Kaggle dataset by applying Naïve Bayes and Random forest classifiers. Furthermore, this study will also attempt to investigate some of the factors of algorithms to demystify their practical application of machine learning models and compare the result with the previous study done by Hong, Kim and Jung (2020)

Initial analysis of the data set including basic statistics

- Dataset is taken from Kaggle dataset based upon the customer survey in the aviation company.
- Original dataset contains 103594 rows and 24 columns, for the training data and 25893 rows for the test data
- Due to the computation reason on the later stage, the dataset was randomly sampled to 5000 and 1250 rows for each from original datasets.
- Dataset has 3 numerical variables, 1 categorical variable and 18 ordinal variables.

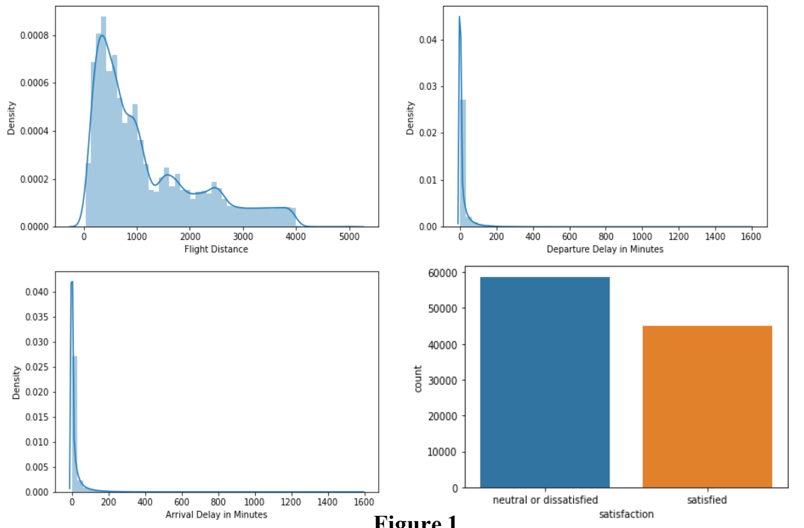

Figure 1

- Figure 1 provides information about the distribution of numerical variables and the target variable. Understandably, as excessive flight delay does not occur on a regular basis, these numerical data have a tendency to be left-skewed. Also, my target column has a binary label (Natural or dissatisfied and satisfied) which is slightly imbalanced.
- Figure 2 illustrates a heatmap by applying Pearson correlation coefficient against the target variable (satisfaction)
- According to the heatmap, overall, there are two distinctive tendencies for the relationship between dependent variable and independent variables.
  1) Unlike the normal expectation, Departure/Arrival information does not have a large impact on the linear independence against satisfaction.
  2) As opposed to Departure/Arrival information, the quality of the flight experience such as class, online boarding and inflight service seem to be correlated with satisfaction.

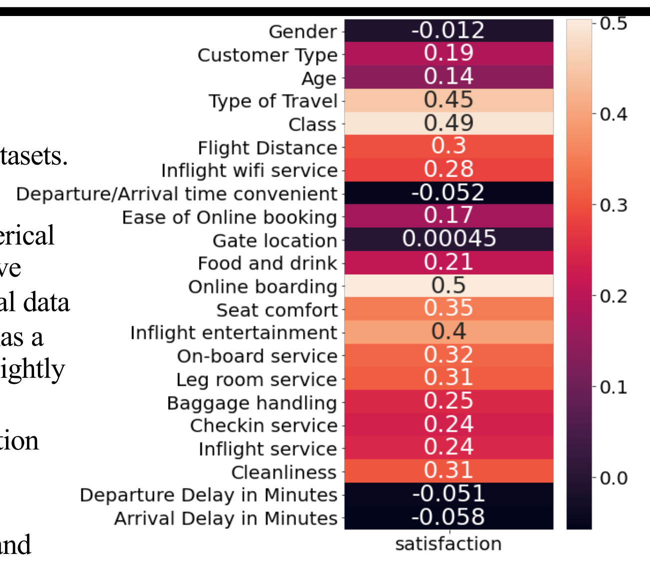

Figure 2

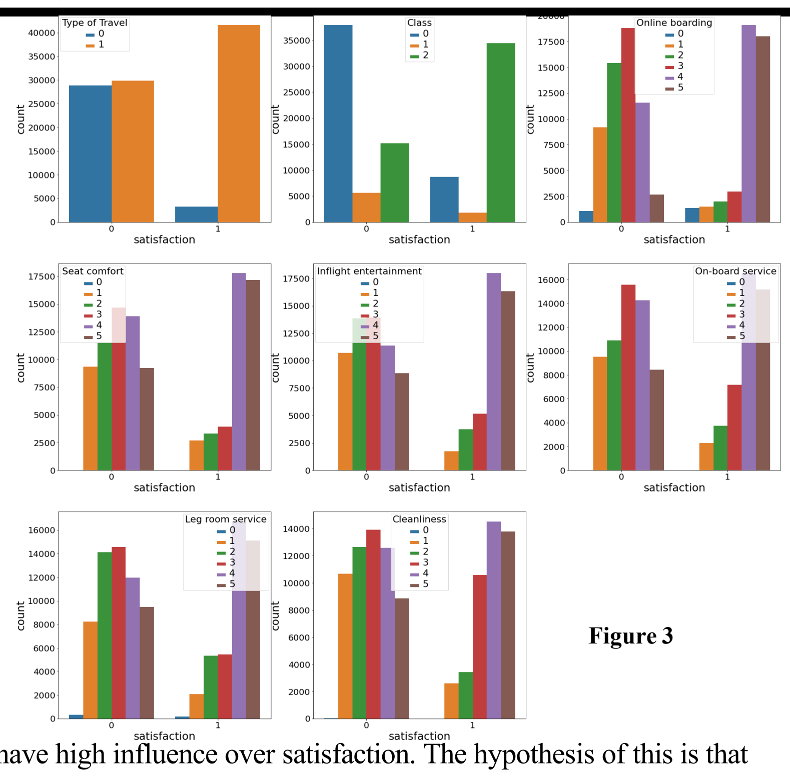

Figure 3

- Figure 3 visualized bar charts between satisfaction and high-correlation columns (>0.3) by different hue for each specification. As can be seen in the heatmap, the type of travel seems to have high influence over satisfaction. The hypothesis of this is that business travel generally can be accompanied by other luxury services which might be contributing to the satisfaction such as the type of seat and on-board service.
- As incorporating un-related model might cause a negative effect on the computational ability and the evaluation metrics, variables showing very weak correlation coefficient such as Departure/Arrival Delay were removed for the modeling.

Brief summary of the two ML models with their pros and cons

Naïve Bayes:

- The idea of Naïve Bayes is based upon the bayes theorem, which is the conditional probability of features for each class (Kaviani and Dhotre, 2017).
- Given prior information, Naïve Bayes computes a posteriori for each observation and output the most probable one as a result (Zhang and Gao, 2011).

Pros:

(1) As Naïve Bayes is non-parametric models, it does not have any assumption for the probability distribution.
(2) Intuitively understandable and fairly easy to implement the model.
(3) Since Naïve Bayes has few hyper-parameters, it is easy to tweak and track the improvement as a visualisation.
(4) Naives Bayes can deal well with small or medium size data set as Naïve Bayes can be flexible to set a prior information.

Cos:

(1) Naïve Bayes assumes each attribute are independent. This assumption could be very problematic as it could be difficult to meet the satisfaction in the real life case scenario.
(2) If some information in the test dataset is not appearing in the train data in the text, Naïve Bayes classifier will output a zero probability for that feature in the test data. This phenomena is called as a Zero-frequency problem. In order to address this, it is essential to count one for each attributes so that each attributes can be recognised at least one time (Garg, 2013).

Random Forest:

- Random Forest is a supervised learning model and widely used in regression and classification tasks.
- Random Forest is a set of combination of decision tree in order to provide a better result in comparison to single decision tree.
- In Random Forest, bootstrapping method is used to train multiple models in parallel, which allows to create a robust model and give the output based upon voting system (Breiman, 2001) and typically trained on a subset of features.

Pros:

(1) As single decision tree has a tendency to overfit easily, random forest can offset this weakness because random forest can use ensemble techniques to combine multiple trees and pick up the strength point from each trees.
(2) Random Forest can provide an output without imputing missing data and has a robust to outliers. Compared with other linear models, it does not require the effort for feature engineering to create a benchmark although there might be a case where feature engineering could produce a better result in the evaluation metrics.
(3) In general, tree model such as Random Forest can visualize the feature importance which gives an insight of each contributions to the predictive model (Scornet, 2020). This could be useful as a way of exploratory data analysis at the beginning.

Cons:

(1) Random Forest has many hyperparameters to tune. Due to this, the difficulty arises when visualisation is required to showcase the result as an interpretation.
(2) Since Random Forest contain many decision trees because of the nature, it tends to be computationally expensive.
(3) In comparison to simple model such as linear regression, model can be black-box, resulting in the complexity of fully understanding the reasoning for the model.

Hypothesis statement

- Random Forest can perform well compared with Naïve Bayes as Random Forest can create more deep trees and would expect to yield better scoring metrics.
- Due to the complexity of the model, Random Forest can take more running time to compute
- Since previous study done by Hong, Kim and Jung (2020) has not conducted hyper-parameter tuning, this model would yield slightly better result.
- That being said, given that the dataset has reasonable amount of information in terms of feature and sample size, both models can be useful for customer satisfaction prediction.

Choice of parameters and experimental results (Naïve Bayes)

Choice of parameters: Distribution Name: Normal, Width: Nan

- Grid search was chosen as a hyper-parameter tuning method. This allows to conduct a complete search over a subset of the hyperparameters space specified by the tuning stage (Liashchynskyi, 2019).
- MaxObjectiveEvaluations is specified as 20 to observe the tendency of hyper-parameters within manageable time constraints.

Experimental results:

- Figure 4 illustrates a confusion matrix for test data as a final output.
- Figure 5 provides information about the sequence of iteration and minimum objective.
- Table 1 shows several evaluation metrics for three different trials. Detailed information is as follows:
  Normal model is Naïve Bayes classifier with default hyper-parameter for train data.
  HP Tuning model is Naïve Bayes classifier with the most optimised hyper-parameter for train data.
  Test model is Naïve Bayes classifier with the most optimised hyper-parameter for test data.

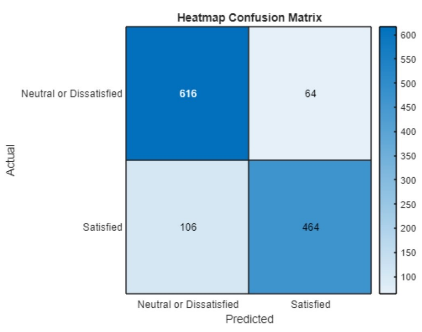

Figure 4

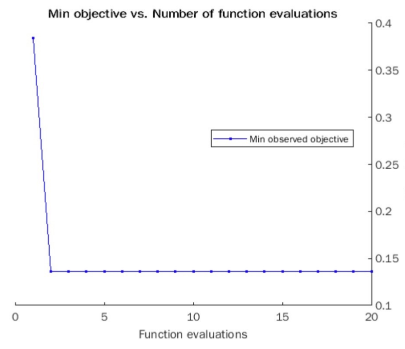

Figure 5

Normal	Metrics	HP Tuning	Test
0.8642	Accuracy	0.8642	0.8640
0.8591	Precision	0.8591	0.8600
0.8638	Recall	0.8638	0.8660
0.8615	F1score	0.8615	0.8630
N/A	Running time	113.9 sec	N/A

Table 1

Description of the choice of training and evaluation methodology

- Due to the computational reason, train data is reduced to 5000 from 103594, and likewise, test data is reduced to 1250 from 25893. Although this is a huge decrease in terms of the sample size, the result would hold a validity as the sample size is expected to be enough to generate an interesting result find a meaningful implication.
- In order to estimate the generalization error (Bengio and Grandvalet, 2004), 10-fold cross-validations is applied in the training dataset.
- Due to the extremely low correlation coefficient, Departure/Arrival delay information were removed for modeling phase.
- In this case study, I will not only compare the result for each model but also for the improvement for before and after hyper-parameter tuning.
- As evaluation metrics, I will investigate several evaluation metrics such as accuracy, precision, recall and F1 score as accuracy might not be able to capture the correctness of prediction from multiple perspectives.

Random Forest:

Choice of parameters: MaxNumSplits: 292 NumVariablesToSample: 9

- Maximal number of decision splits and Number of predictors to select at random for each split are used as a tuning hyper-parameter.
- MaxObjectiveEvaluations is specified as 25 to fully observe a learning trend.
- Likewise, grid-search was applied to tune the hyper-parameter.
- Surrogate was specified as 'off'. This could potentially improve the score if feature contains missing value. However, since there is no missing value in the modified dataset, this specification has been made.
- Similarly, each figure and table show the result for test data and the learning curve for minimising objectives against iteration of grid-serach
- Figure 8 visualises the three-dimensional surface plot for optimizing a hyper-parameter. Each light red point corresponds to the examined hyper-parameter during the grid search. Red point refers to the global optima for the grid search shown above as a choice of parameters.

Normal	Metrics	HP Tuning	Test
0.9388	Accuracy	0.9434	0.9456
0.9361	Precision	0.9406	0.9436
0.9408	Recall	0.9442	0.9470
0.9391	F1score	0.9424	0.9453
N/A	Running time	369.19 sec	N/A

Table 2

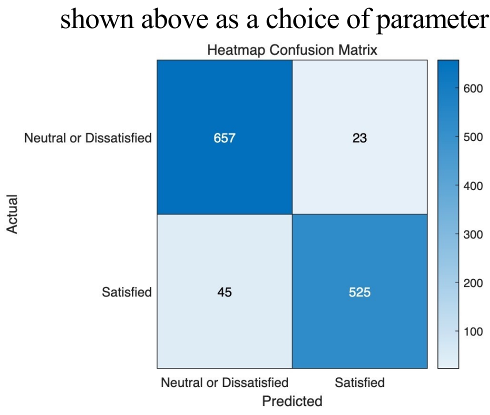

Figure 6

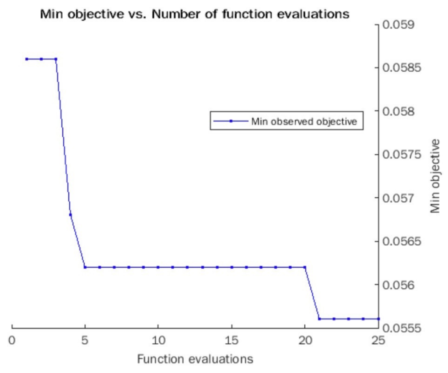

Figure 7

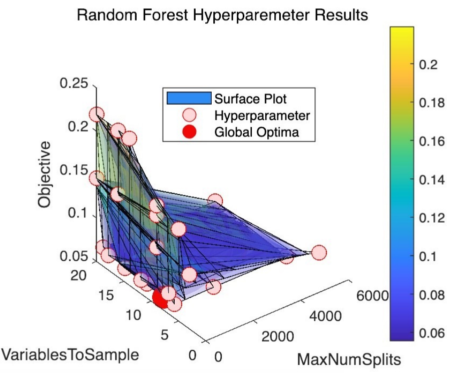

Figure 8

Analysis and critical evaluation of results

- In Naïve Bayes, one possible assumption for Normal distribution as opposed to Kernel Density distribution is chosen is that most of the variables used in the prediction is normally distributed ordinal variables such as online-boarding. Hence, Normal Distribution might be able to capture the trait easily.
- Since best hyper parameter for Naïve Bayes is normal distribution, width which is only used for kernel density distribution was shown as 'Nan'. Compared with non-hyper parameter tuning model, the result was exactly the same. This is due to the fact normal distribution is default for Naïve Bayes. Unlike the common understanding, hyper parameter tuning does not contribute to the improvement of scoring metrics in this case study.
- In Random Forest model, the number of false negatives are larger than false positive. This tendency can be observed in precision (0.9361) and recall score (0.9408) from Table 2. One possible reason for this is the imbalanced data for the target feature. Although there is no significant imbalance compared with medical dataset, this might have an influence on giving a slightly lower score on precision.
- In Random Forest, even after drastically decreasing the sample size, Random Forest takes roughly 6 minutes to complete hyper-parameter tuning with only two parameters. In order to yield the best result, training the result with many hyper parameter plays a central role in the improvement of the model. This implies that there is a problematic trade-off relationship between time constraint and the improvement of scoring metrics. Therefore, importance should be placed upon the deployment of the real business case. For instance, in the deployment stage, taking a financial return against the time constraint into consideration could be key factor.
- As can be described in the initial hypothesis, although Random Forest performs better than every evaluation metrics, both models can generally perform in the practical level and no particular overfitting and underfitting are observed.
- As a nature of Random Forest, low variance can be seen due to Random Forest uses bagging technique which is to combine several weak learners and 10-fold cross-validation. Furthermore, the choice of Max number of decision split during a hyper parameter tuning also might contribute to low variance as this can prevent overfitting from the model.
- When I compared Normal model with HP tuning model for train data, the latter does not necessary require cross-validation as HP tuning model is already optimised. However, in order to conduct a fair comparison between them, cross-validation was applied.
- Compared with 95.4% accuracy for Random Forest in the previous study done by Hong, Kim and Jung (2020), the result (0.9456) is slightly lower than the previous research. Given that the dramatical decrease for the number of data (from 103594 to 5000) in this case study, it seems to be true that this model is also robust and the result holds validity.

Lessons learned and future work

- In the practical application, decision for the trade-off relationship between the constraint such as computationally ability, time, the objective for the predictive model and improvement of the score has to be determined carefully.
- Hyper parameter tuning might not always improve the score. Although it provides the room for the improvement, that could be a small amount. Therefore, it could be better to spend more time on feature engineering rather than tuning the hyper parameter.
- As grid-search attempts to compute every possible combinations for hyper-parameter, random search or bayesian optimization could be more efficient measure to optimize hyper parameters.
- As a general concern, machine learning tends to be black-box. Hence, along with the predictive model, explainable AI technique such as permutation feature importance or alpha value might be valuable to demystify the contribution to features.
- In addition, implementing a more advanced model such as gradient boosting decision tree with full set of data can provide a better result.
- Due to the slight imbalanced data for target model, imbalanced technique such as SMOTE or stratified cross-validation could be worthwhile to generalise the model to the application for future work.

References

[1] Hong, S., Kim, B. and Jung, Y., 2020. Correlation Analysis of Airline Customer Satisfaction using Random Forest with Deep Neural Network and Support Vector Machine Model. *International Journal of Internet, Broadcasting and Communication*, 12(4),26-32

[2] Kaviani, P. and Dhotre, S., 2017. SHORT SURVEY ON NAIVE BAYES ALGORITHM. *International Journal of Advance Engineering and Research Development*, 4(11),pp.607-611.

[3] Zhang, W. and Gao, F., 2011. An Improvement to Naive Bayes for Text Classification. *Procedia Engineering*, 15, pp.2160-2164.

[4] Garg, B., 2013. *DESIGN AND DEVELOPMENT OF NAÏVE BAYES CLASSIFIER*. [online] Library.ndsu.edu.
Available at: <https://library.ndsu.edu/ir/bitstream/handle/10365/23048/Garg_Design%20and%20Development%20of%20Naive%20Bayes%20Classifier.pdf?sequence=1&isAllowed=y> [Accessed 7 December 2021].

[5] Breiman, L., 2001. *RANDOM FORESTS*. [online] Stat.berkeley.edu. Available at: <https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf> [Accessed 7 December 2021].

[6] Scornet, E., 2020. *Trees, forests, and impurity-based variable importance*. [online] Hal.archives-ouvertes.fr. Available at: <https://hal.archives-ouvertes.fr/hal-02436169v2/document> [Accessed 7 December 2021].

[7] Bengio, Y. and Grandvalet, Y., 2004. No Unbiased Estimator of the Variance of K-Fold Cross-Validation. *Journal of Machine Learning Research*, [online] 5, pp.1089-11-5. Available at: <https://www.jmlr.org/papers/volume5/grandvalet04a/grandvalet04a.pdf> [Accessed 7 December 2021].

[8] Liashchynskyi, P., 2019. *Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS*. [online] Available at: <https://arxiv.org/pdf/1912.06059.pdf > [Accessed 7 December 2021]