Predicting customer satisfaction (Invistico Airlines)

Abstract

To solve the challenge of customer retention, the aim of this project was to use passenger feedback data to construct machine learning (ML) systems to predict passenger satisfaction and inform the airline which services are the most important for passenger satisfaction. A Random Forest Classifier and XGB Classifier were each tuned, trained, and applied to predict passenger satisfaction; ultimately, the models were both successful because of their high recall metric scores (~ 93% on the test dataset) that were significantly greater than the baseline model's and comparable to optimised solutions for similar classification tasks. From the models' feature importance scores, the most important factors affecting passenger satisfaction were in-flight entertainment, seat comfort, and ease of online booking.

Introduction

A major challenge facing corporations is customer retention, particularly for airlines that operate within an extremely competitive industry with tight profit margins meaning that customer churn is especially damaging. The aim of this project was to, for The purpose of this project was to use an anonymous airline's customer feedback data to:

1. Construct machine learning (ML) systems capable of predicting passenger satisfaction with high precision, such that loyalty/benefit schemes can be targeted to specific passengers with a higher probability of dissatisfaction.

2. Provide recommendations regarding the most important aspects of an airline's service, which impact upon customer satisfaction, that airlines should invest in.

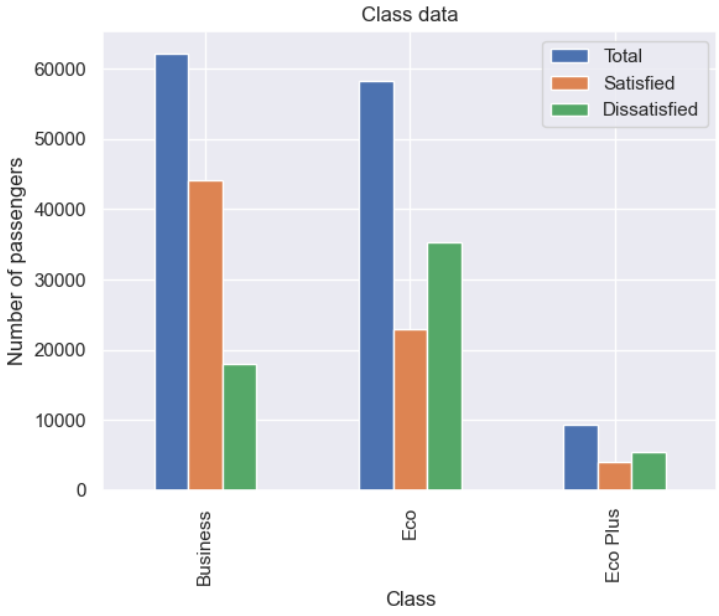
In this project, passenger feedback data from an anonymous airline (labelled 'Invistico') applied for the training and testing of a Random Forest Classifier and an XGB Classifier to determine which provided superior performance, assessed by considering recall and accuracy.

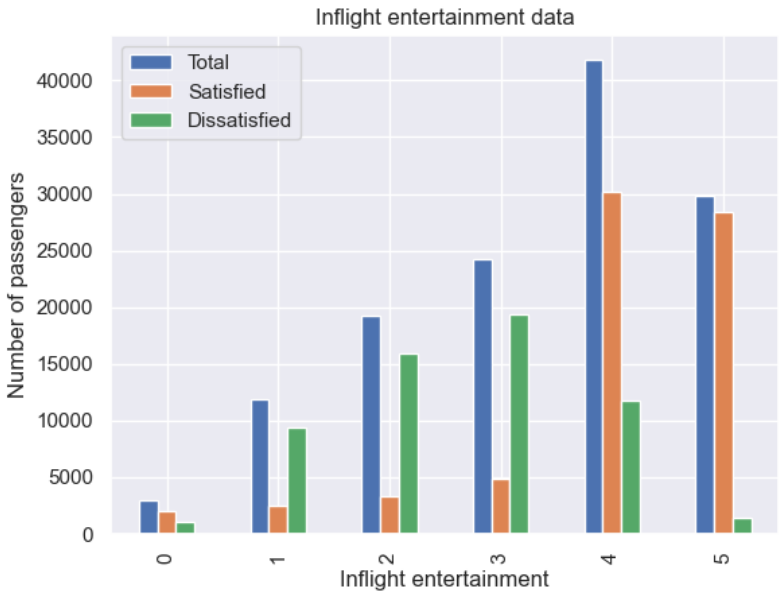
Methodology

Dataset

The data set used in this project can be found on Kaggle. The data has been provided by an anonymous airline under a fabricated name ('Invistico Airlines') and consists of passenger feedback and flight details. There are 129,880 samples and 23 features. The majority of the features contain passenger feedback data regarding the quality of various airline services/amenities rated on a scale of 0 to 5 (note that 0 represents N/A, where the passenger declined to provide a rating). The rest of the features consist of numerical and categorical details relating to passenger data (e.g. gender, age etc) and flight data (e.g. flight distance, departure/arrival delay times etc).

The primary focus of the exploratory data analysis (EDA) was to uncover the features that were closely correlated with satisfaction. Because satisfaction, as well as most other features, were categorical, bar charts were generated for each feature denoting the relationship between the satisfaction and the given feature. Some examples are displayed below:





A few key observations were noted:

1. There were a similar number of male and female passengers, but the satisfaction rate was significantly higher for female passengers.

2. Approximately five times more loyal passengers than disloyal; disloyal have satisfaction rate far lower than 0.5, whereas loyal customers have satisfaction considerably higher than 0.5. This confirms that customer loyalty is directly related to overall satisfaction with services provided.

3. Similar number of business class and economy class passengers; far fewer economy plus passengers. Overall satisfaction declines with class; business class passengers have satisfaction rate significantly higher than 0.5, whereas economy plus and economy passengers have satisfaction less than 0.5. Thus, future investment should be focused on improving services for these classes.

4. As expected for passenger feedback features: the higher the rating, the higher the probability of passenger satisfaction. Overall satisfaction rate surpasses 0.5 when the service rating is equal to or greater than 3 or 4.

Full details and observations from the EDA can be found in the Jupyter Notebook. Evidently, identifying correlations between satisfaction and the other features was challenging since satisfaction and most others are categorical features thus relied upon human judgement. The interconnections between different features would require significant time to investigate. Therefore, training and applying machine learning models, from which predictions and feature importance rankings can be extracted, is a sensible method to proceed with.

Baseline Model

The baseline model was constructed using a Dummy Classifier with the 'most frequent' strategy applied to predict satisfaction since there is minimal class imbalance and so the strategy will provide a valid baseline performance to compare the ML models performance.

Algorithms

The two ML models applied were the Random Forest Classifier and XGB Classifier. These are both ensemble methods, where both ensembles apply decision tree classifiers to make predictions. Decision tree classifiers choose the optimal feature, considering information gain, by which to split the dataset. This process is repeated until a defined limit is reached (e.g. maximum tree depth can be set during initialisation). The samples in each subset (i.e. 'leaf')

are assigned a class label dependent upon the majority class within the subset.

The Random Forest Classifier involves recursively applying a pre-defined number decision trees on a random subset of the training data (i.e. bootstrap sample). The decision trees are independent of one another and are all used to classify the target feature; the final result is whichever value is predicted by the majority of trees.

The XGB Classifier involves initially defining a decision tree with a 'dummy' prediction (e.g. constant value for all values of target feature). The gradient of the loss function for the model's prediction is calculated for each sample and a new decision tree is fit with its structure determined via an algorithm that optimises the objective function. The new tree is applied to make predictions and the cycle of optimisation for the creation of improved trees is repeated for a preset number of iterations.

The Random Forest Classifier and XGB Classifier were chosen since they are industry-standard tools for classification tasks. These ensemble methods are superior to individual decision trees since they are less prone to overfitting and more robust to noise since they aggregate predictions from multiple decision trees and so can handle complex feature relationships thus making superior predictions.

Metrics

Two metrics were used to assess model performance. The priority for the airline is to correctly identify which passengers are dissatisfied; thus, the primary metric was recall, which measures the proportion of actual positive instances (i.e. dissatisfaction cases) that were correctly labelled. This can be summarised in the following formula:

$$\text{Recall} = TP / (TP + FN)$$

Where TP is the number of correctly-identified dissatisfied passengers and FN is the number of incorrectly-identified dissatisfied passengers. The target value for recall is 1.

The secondary metric applied was accuracy, which measures the proportion of all passengers whose satisfaction was correctly predicted. The following formula illustrates this:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Where TN is the number of correctly-identified satisfied passengers and FP is the number of incorrectly-identified satisfied passengers. The purpose of the accuracy metric was to supplement the assessment of recall by providing an indication of overall model accuracy.

Experiments

The performance of untuned ML models and the baseline model were compared to determine whether the untuned ML models had greater recall/ accuracy than the baseline model so that there is cause to proceed with fine-tuning the models. The ML models had significantly greater metric scores than the baseline, so fine-tuning the models was justified.

For the Random Forest Classifier, the hyper-parameter search spaces were as follows:

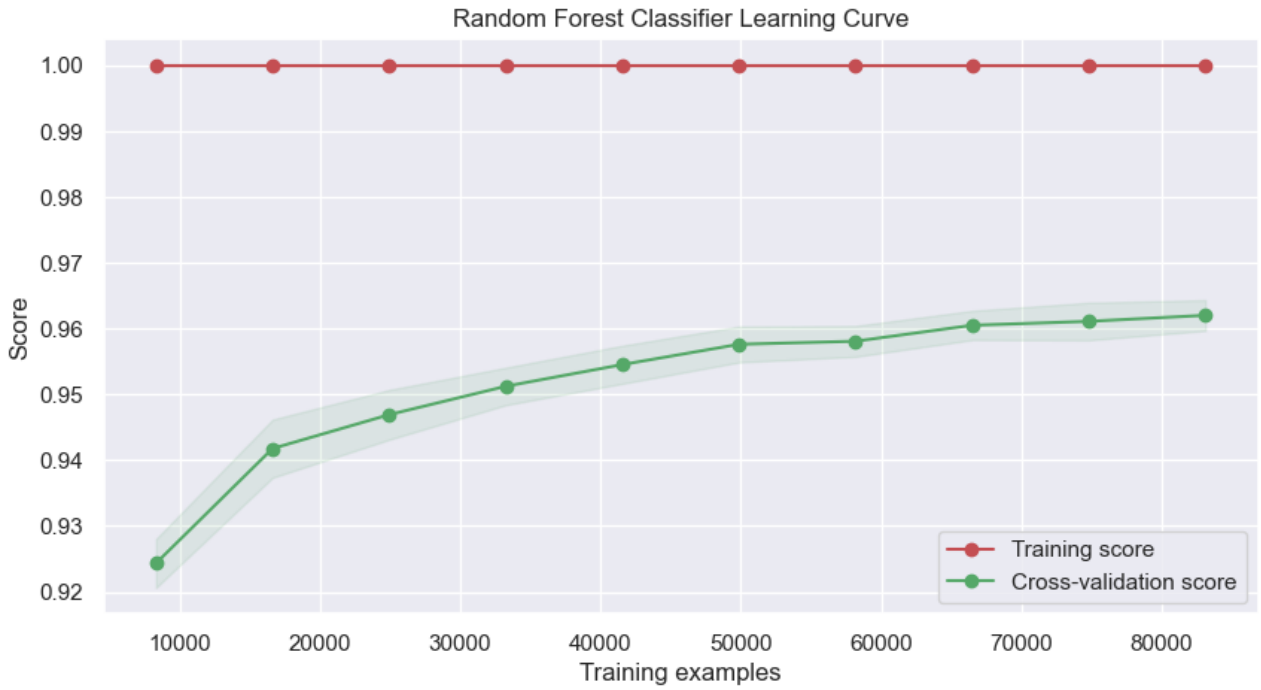
- max_depth (maximum number of layers for each tree) - 1 to 50
- n_estimators (number of trees in the model) - 100 to 500
- min_samples_split (minimum number of samples in leaf to split) - 2 to 100
- min_samples_leaf (minimum number of samples required to form leaf) - 1 to 100

For the XGB Classifier, the hyper-parameter search spaces were:

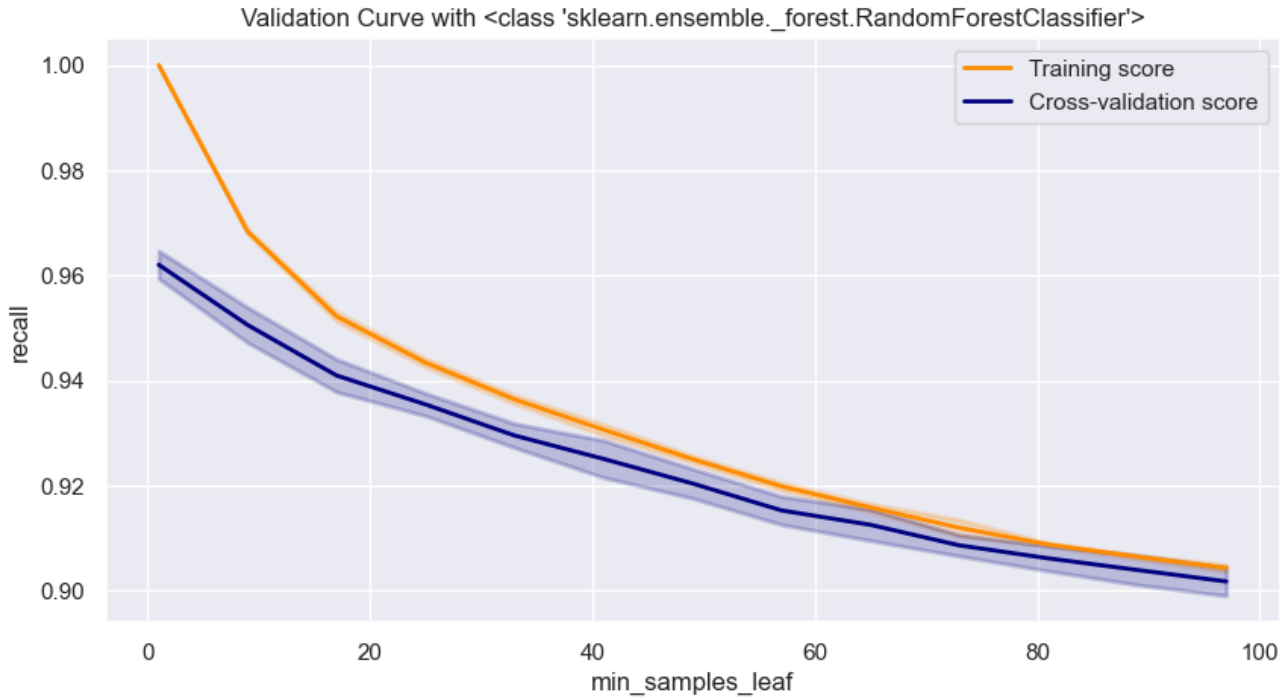
- max_depth (maximum number of layers for each tree) - 1 to 20
- n_estimators (number of 'boosting rounds') - 100 to 400
- learning_rate (step size during each boosting round) - 0.01 to 0.3
- subsample (fraction of total samples applied for tree training) - 0.1 to 1

Bayes Search Cross-Validation was applied to find an optimal model since the search spaces for both models were quite large, which suits the optimisation approach of the Bayesian Search since fewer iterations than Randomised or Grid Search would be needed to obtain sufficient results (i.e. more time efficient for this case).

Learning curves for each model and validation curves for each hyper-parameter was generated to check for and amend under-fitting and over-fitting:



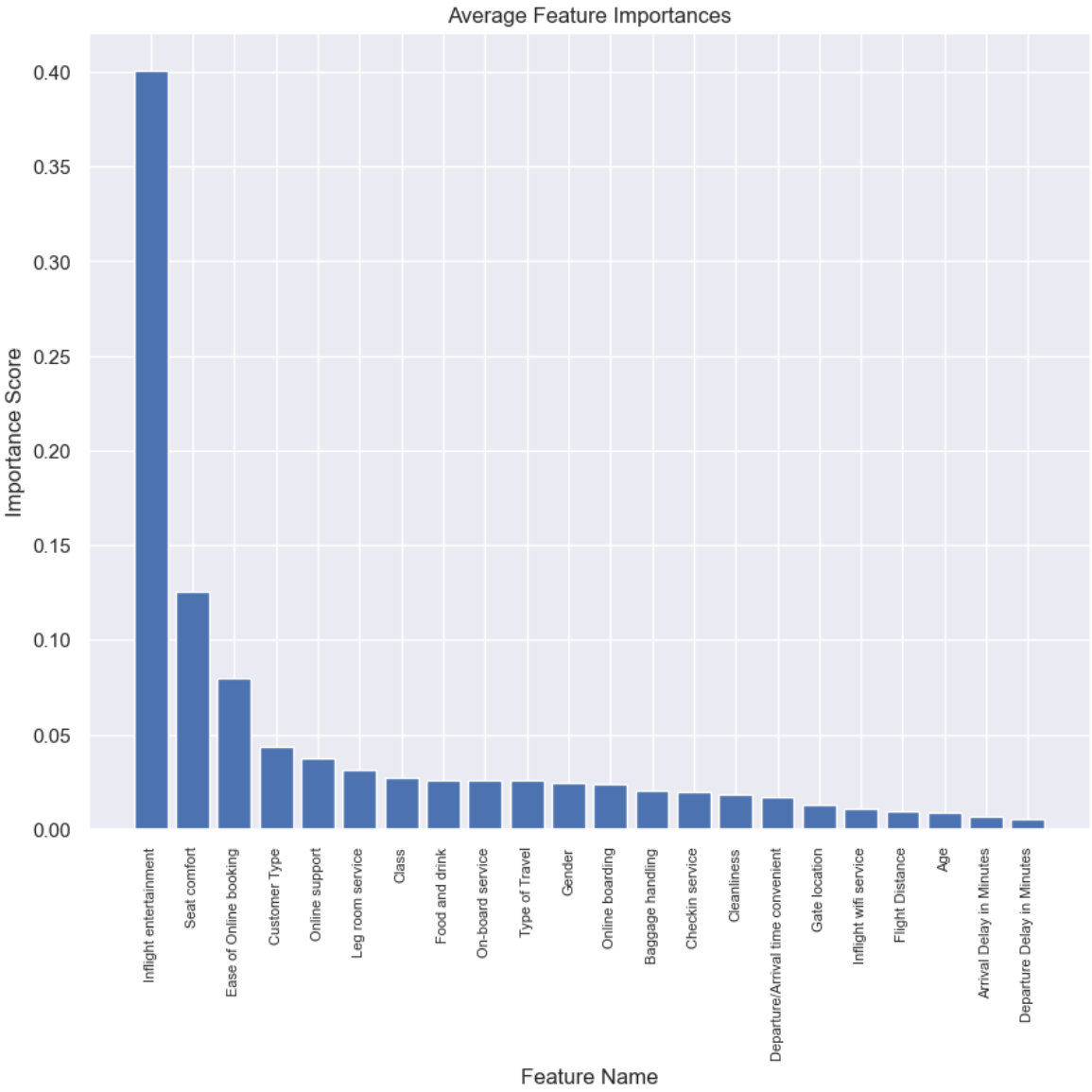
In the Random Forest's learning curve, the recall score of the model applied to cross-validation data rises but eventually plateaus, indicating that more samples would have negligible impact upon model quality. However, the score for the training data does not decrease from 1, strongly indicating over-fitting that needs to be corrected by adjusting the model's parameters. For example:



The current optimal model's value for min_samples_leaf is 1. However, as the value of min_samples_leaf decreases from 15 towards 1, the recall score on the training data diverges away from the cross-validation score suggesting over-fitting. Thus, min_samples_leaf was set to 15 to prevent over-fitting. A similar process was followed for all other hyper-parameter validation curves.

Results and Analysis

Score		Random Forest Classifier	XGB Classifier	Dummy Classifier
Recall	Training	0.9343	0.9347	0.4997
	Test	0.9314	0.9328	0.4934
Accuracy	Training	0.9410	0.9407	0.5025
	Test	0.9372	0.9389	0.4995



Both the Random Forest Classifier and XGB Classifier have similar metric scores when applied to the training and test datasets. The recall scores are of most interest, as explained before, and are high (~ 93%) and comparable with results from ML models applied to similar classification tasks.

The feature importance scores are an attribute of trained models. As each model applied utilises a different ensemble technique, the respective importances for each feature differ slightly; thus, the feature importances were averaged across the two models and are displayed in the following bar chart on the previous page.

The features are ranked in terms of importance to passenger satisfaction. The top three most import features, that the airline should prioritise, are inflight entertainment, seat comfort, and ease of online booking. Feature importance declines exponentially; if there was significantly more features in the dataset whose noise was affecting model performance, this feature importance data could be utilised to remove relatively unimportant features below a defined threshold.

Conclusion

In this project, ML models were successfully trained and applied to predict customer (dis)satisfaction with high performance. The recommendation is to use the XGB Classifier since its performance was marginally superior to the Random Forest Classifier, the gradient descent algorithm results in greater scalability and fasting training time, and is known to be more efficient at handling non-linear, complex relationships between features present in this dataset.

Nevertheless, there are numerous avenues for future exploration: how does higher dimensionality (i.e. inclusion of more features such as flight destination or additional customer feedback fields) impact on model performance? In this case, would less common classification methods (e.g. Support Vector Machine) obtain better results due to differences in ability to handle higher dimensional data with more complex relationships. These are thoughts to ponder in future classification scenarios if these conditions are present.