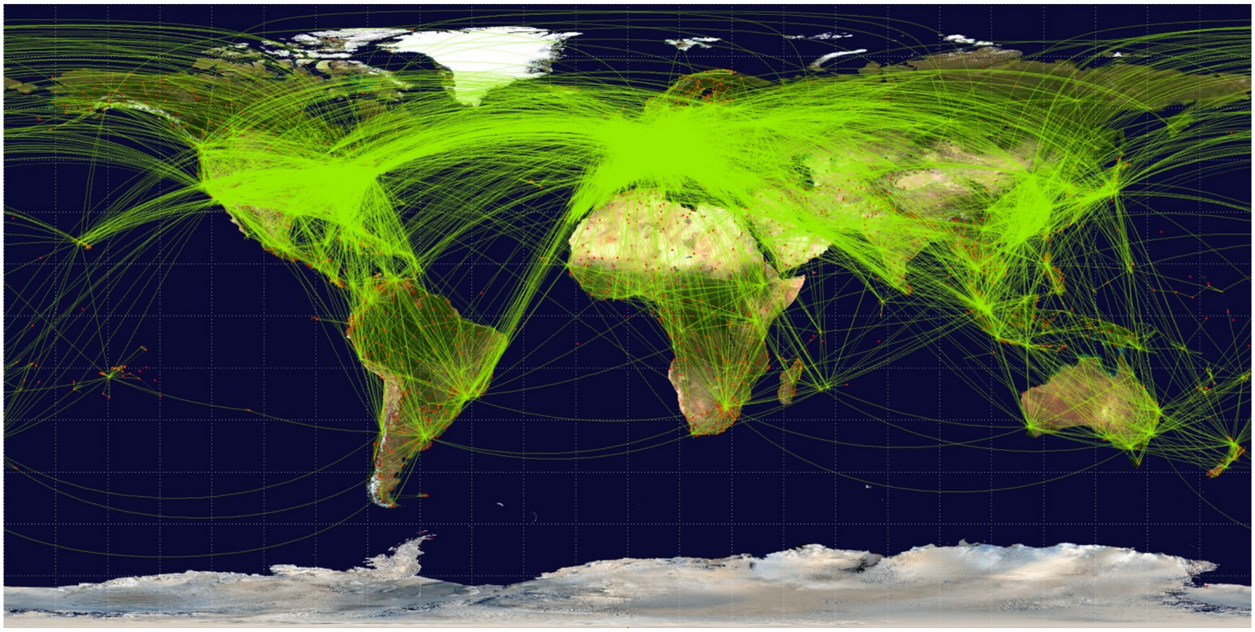


Image: wikimedia.org

Invistico Airlines Project Report: 2nd case

For 1st and 2nd case code: https://github.com/izsolnay/Invistico_Airlines_Python

Objectives

2nd case

Invistico Airlines is interested in predicting whether a future customer would be satisfied given previous customer feedback about their total flight experience. The airline would like the construction and evaluation of a model that can accomplish this goal. They are also interested in knowing which features of the total flight experience are most important to customer satisfaction.

Data

The data is sample size of survey responses from 129,880 customers. It includes data points such as satisfaction, class, flight distance, and inflight entertainment, among others. It is an in-house product.

Deliverables

Models: binomial logistic regression, decision tree, random forest, XGBoost.

(Since this project is a portfolio addition, the models are predetermined. Since this is a labelled data set, all models qualify as supervised.) Python Jupyter notebooks were used for all coding.

2nd case: Tree-based classification machine learning models (2nd notebook)

In the second case, the airline is interested in predicting whether a future customer would be satisfied with their flight experience given previous customer feedback. They are also interested in knowing which features of their flight experience are most predictive of that satisfaction. Decision trees are classification models that can handle all of the types of variables this data set contains (categorical, discrete, and continuous). They make predictions for a target variable based on multiple variables (features). Because decision trees return a weighted map of possible outcomes to any series of related choices, the airline can use these findings to allot their available resources.

For this second case 3 tree-based models are trained, tuned, evaluated, and compared for the most predictive power. The most predictive will be tested on the unseen test data.

1. Simple decision tree: can handle collinearity and has no assumptions about data distribution. This works well for this data set, since its variables have highly uneven distributions, scales, and data types.
2. Random Forest: these models leverage randomness by combining the results of many models to help make more reliable final predictions. These predictions have less bias, and lower variance than other standalone models. They are also very scalable (thus can handle outliers easily, which is also a factor in this data set) and less susceptible to overfitting.
3. Gradient Boosting (XGBoost): these models have a high accuracy rate, work well with missing data (which this data set has), are scalable, and can provide a weighted feature importance chart. However, as they are considered black-boxes, their steps cannot be seen (the trees cannot be produced).

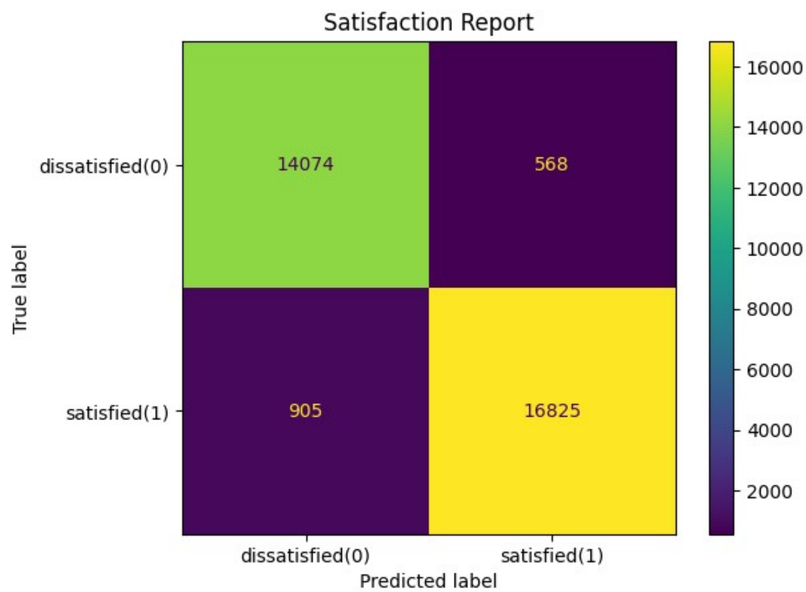
The objective is to utilize the most predictive decision tree model to forecast whether or not a customer will be satisfied(y) with their flight experience based on the multiple features (variables) in the data set(Xs) and discover which variables are most predictive of that outcome.

The process

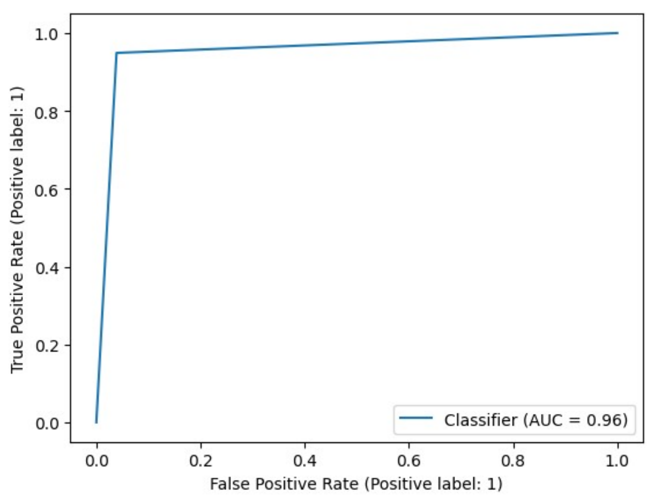
First a baseline decision tree model was built and tested to get a feel for predictability. The model scored very well with all metrics above 90%, indicating that further tuning would likely result in a high quality model. As expected, the succeeding tuned decision tree scored several points higher. Next, a random forest model was built. This had even better metrics, which were confirmed using a split validation data set. Finally, an XGBoost model was built. Surprisingly, this model had the worse performance of all of the tree-based models (even lower than the baseline decision tree). The winning model was the random forest.

Results on test data

From the 32,372 in the test data, 17,393 were satisfied. Of those, the model captured 16,825 satisfied customers – missing a mere 568 satisfied customers), thereby achieving a ~95% recall score. The test data set also included 14979 dissatisfied customers. Of these, the model captured 14,074 – missing 905 satisfied customers. Thus, it was slightly better at predicting who would be satisfied, which was exactly what it was designed to do.



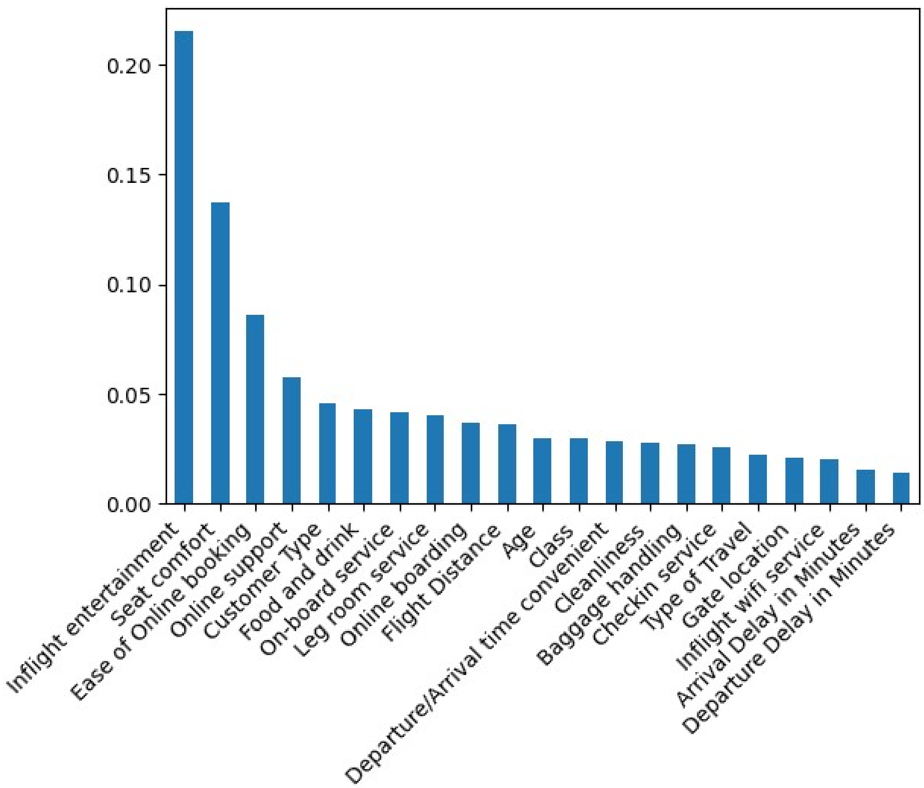
The model's precision of ~97% (with a f1 of ~96%), means that it has an equally high ability to identify satisfied or unsatisfied customers and an equally low chance of falsely identifying satisfied or unsatisfied customers (Type I and Type II errors). Further demonstrating this model's efficacy, it achieved an AUC score of ~96%, which suggests a near perfect performance over all possible classification thresholds in distinguishing between positive and negative classes.



Because the model scored so highly in all metrics, it is difficult to suggest further optimization. Optimizing either precision or recall scores, would likely result in either more false positives, where it would attempt to reduce the occurrences of incorrectly classifying people as satisfied when they are dissatisfied (Type I errors) or more false-negatives where more people would be classified as dissatisfied, when then are satisfied (Type II errors).

Most predictive features

The second objective of this random forest model was to uncover which features of a customer's airline experience were most predictive of their satisfaction. The most predictive feature in customer satisfaction as ascertained by the model was Inflight entertainment, with Seat comfort and Ease of online booking also being somewhat predictive. Inflight wifi service was not the least predictive variable as it had been in non-winning models; however, perhaps surprisingly, delays either in arrival or departure had the lowest predictive effect.



The extreme success of the model indicates that the airline should focus attention on their inflight entertainment service, either leaving it as it is or continuing to improve it based on customer feedback. Further resources could be invested in maintaining or improving seat comfort and a smooth booking experience. Finally, as the model scored an ~96% on accuracy, this is far better predictor than the ~55% satisfaction rating for the overall dataset.