AIRLINE CUSTOMER SATISFACTION ANALYSIS & PREDICTION

Submitted in partial fulfillment of achievement criteria for UT McCombs PG-DSBA Program

SUBMITTED BY

Eric Green

PG-DSBA 2020-2021 Oct

Acknowledgements

Acknowledgement is given with a deep appreciation for the passion, disciplined efforts and rigorous standards implemented by McCombs/Great Learning lecturers, mentors, program managers and support staff of the PG-DSBA Program. As a cohort learner, the PG-DSBA program has been an exceptionally high-quality learning experience which has provided me with, not only a powerful set of analytical and machine learning techniques to employ in industry, but also with improved problem solving and experimental acumen and sophisticated mental models for designing and managing modern business agility and transformation strategies in the context of modern data technologies.

Within my own personal context and journey through life and career, I am confident that the knowledge, techniques and rigor I have acquired through the PG-DSBA program will serve as a catalyst for leading-edge pursuits in operational decision support and cognitive automation technology development within my role as Head of Engineering Operations at Adobe as well as for disruptive entrepreneurial endeavors I envision undertaking in the later stages of my life. Thank you, Great Learning.

Date: August 21, 2021 Eric Green

Austin, Texas

TABLE OF CONTENTS

EXECUTIVE SUMMARY

In the 21st Century aviation and airline industry specifically, investment in data technology and data-driven decision-making must be viewed by business leaders, not only as a core strategic aim for the company, but also as fundamental to profitability and survival. To achieve “Apex Predator” status within the market ecosystem, it is imperative for airline companies to effectively, efficiently and intelligently manage data collection, data storage, data retrieval, analytical processing and point of use data delivery to enable action in real-time on feedback and insights to manage business risk and drive deep customer loyalty. The most fit-for-purpose industry players will transform their data directly into business agility and revenue – Data is the new gold and machine learning is the gold excavator.

Falcon Airlines aims to maximize marketing performance, profitability and cost reduction by leveraging data collected on passengers of its airline flights. Falcon Airlines is targeting broad optimization across digital customer services, airport services and infight services to its customers through analysis and modeling of customer attribute and survey data to understand the drivers of customer satisfaction. Through the insights gained analyzing customer data, specific business strategy recommendations are put forth to Falcon Airlines senior leadership.

The report which follows is intended to provide a tangible and implementable strategic improvement path for Falcon Airline’s senior executives in the areas of marketing, operations and infight service delivery. Three separate data sets have been provided which describe information about customers, flights taken, and survey (sentiment) responses.

Through iterative exploratory data analysis and data preprocessing, competing supervised machine learning classification models are generated to identify and map associations between independent variables and the target variable (Satisfaction) to enable predicting which customers are satisfied vs. not satisfied in order to customize marketing and services with maximum efficacy.

Ensemble classification models are generated and testing with different input variants to understand the effects of data preprocessing treatments. The initial stage of model building, and testing serves as a benchmarking to identify which input variants and base algorithms should be selected for downstream model tuning and ensemble model stacking. Subsequent parametric analysis of satisfaction drivers is performed on the stacked combination of models with the objective to achieve the highest Precision score on out-of-sample data with the smallest delta between training and testing data to minimize model overfitting (misclassifications on future data). Once the best model sequence is determined, specific business recommendations are made from the insights generated.

ANALYTICAL LEARNING LIFECYCLE

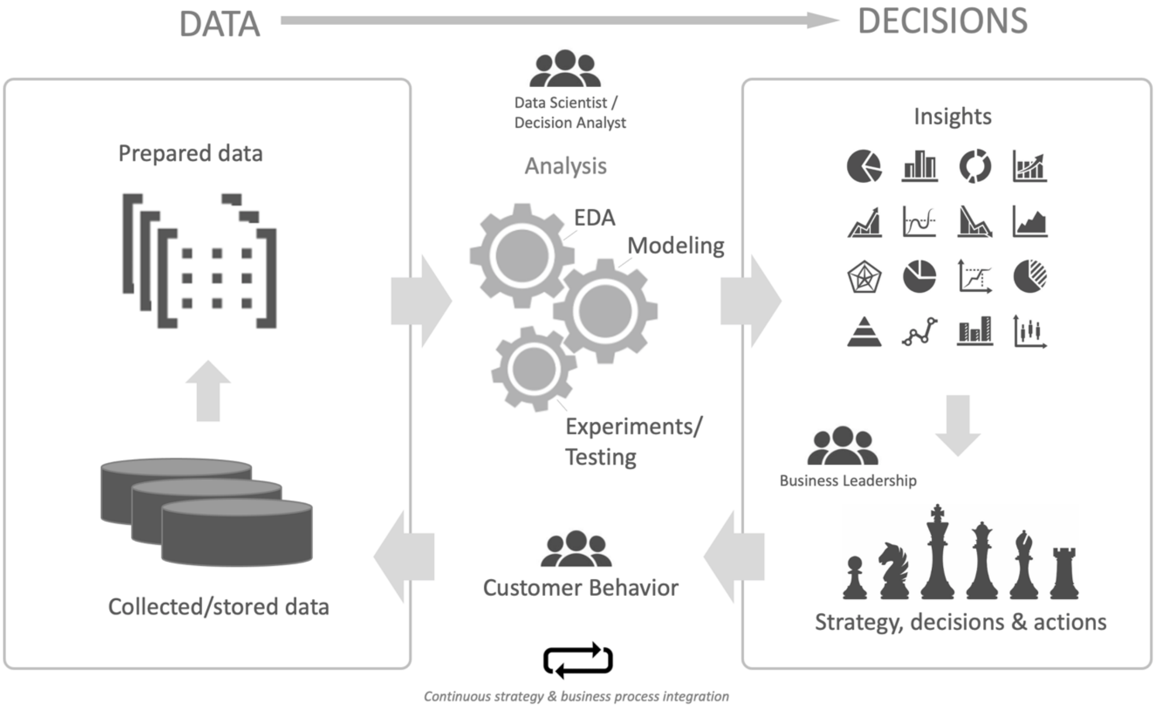


Figure 1: Modern analytical learning & strategic decision-making

INTRODUCTION

Due to the COVID-19 global pandemic and the associated economic chaos and disruption to airline operations globally that have resulted from it, it is more important than ever for customer and survey data to be leveraged to model and understand customer satisfaction and what drives it. From this understanding and strategic awareness, we present a basis for optimizing the value of investments undertaken by Falcon Airlines to drive future growth and performance. The following study and analysis aim to address the following business opportunity landscape.

The below strategic business areas are considered in this report:

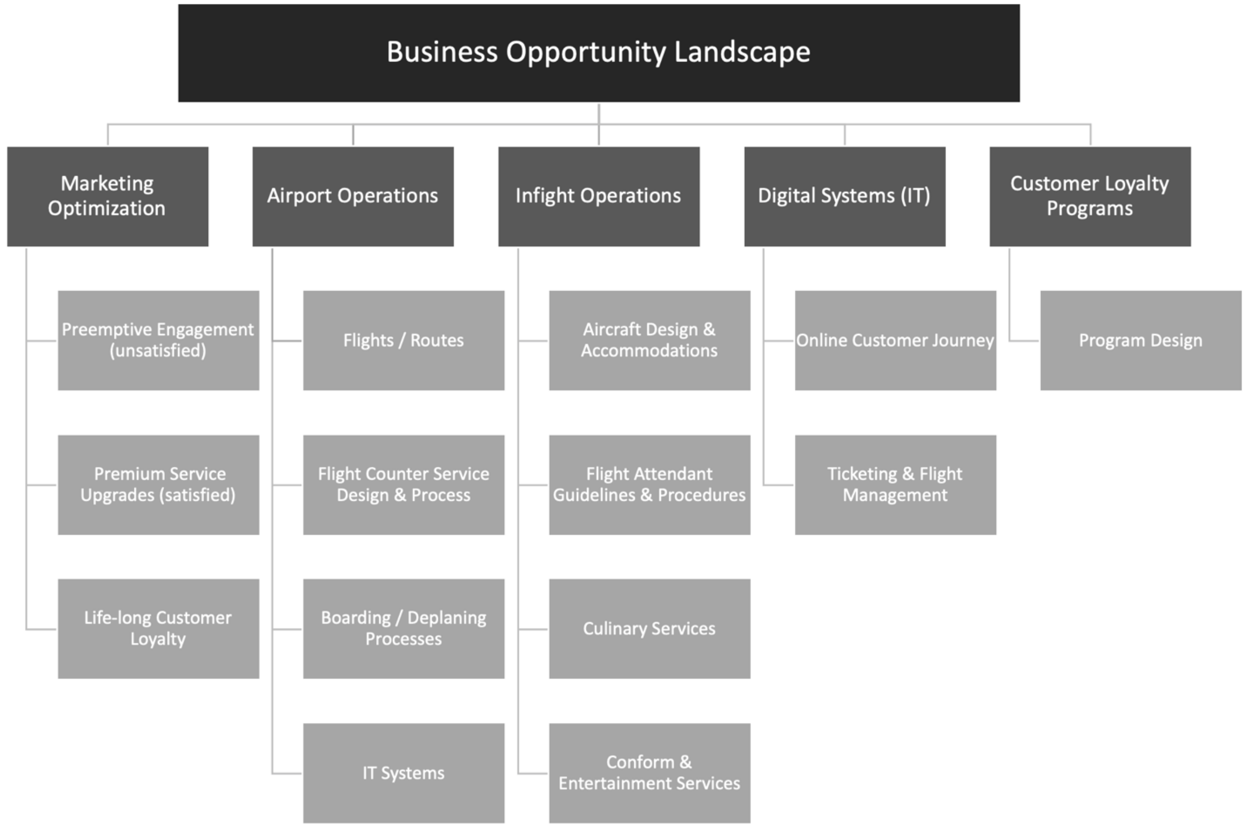

Business Opportunity Landscape
Marketing Optimization
Airport Operations
Infight Operations
Digital Systems (IT)
Customer Loyalty Programs
Preemptive Engagement (unsatisfied)
Premium Service Upgrades (satisfied)
Life-long Customer Loyalty
Flights / Routes
Flight Counter Service Design & Process
Boarding / Deplaning Processes
IT Systems
Aircraft Design & Accommodations
Flight Attendant Guidelines & Procedures
Culinary Services
Conform & Entertainment Services
Online Customer Journey
Ticketing & Flight Management
Program Design

Figure 2: Airline strategic improvement areas considered

By locating patterns in the customer attribute and survey data and correlating these patterns with appropriate business areas above, it is possible to radically improvement business agility and fitness-for-purpose of the airline through targeted investments across marketing and operations in allocating smart (measured) budget ratios and improvement projects in specific areas.

Supervised machine learning-based classification provides a powerful tool to unlock and understand customer satisfaction and its drivers.

Problem Statement

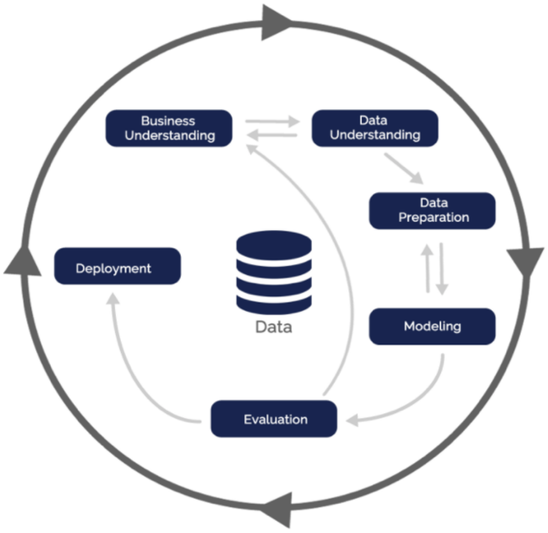
1. The business lacks a quantitative understanding of what drives customer satisfaction and how to efficiently market to customers likely to leave to prevent this from happening.
2. The business lacks technical methods and business process integration needed to realize a data-driven decision-making management framework to sustain and grow airline profits in a volatile global future
3. The business lacks a rapid application development capability in data technologies and artificial intelligence to solve problems within sales, marketing and operations.

Objectives & Scope

1. To build a highly precise and statistically robust classification model to predict customer satisfaction with limited data about the customers.
2. To identify the drivers of customer satisfaction and subsequently recommend improvements across business functions to increase the ratio of satisfied vs. unsatisfied customers.
3. The scope of analysis and business recommendations is broad as it comprehends airline marketing, operations, digital/IT systems and customer loyalty programs.

Analytical Approach

We employ a generalized variant of the Cross-Industry Standard Process for Data Mining (CRISP-DM) *meta model* as a realization of the analytical learning concept illustrated in the Executive Summary. It should be noted that applying the CRISP-DM model below in practice is a heavily iterative experimental process and is not sequential or linear. Cycling through the outer and inner loops will occur an indeterminant number of times for a given business problem to be solved.



Source: blog.magrathealabs.com

Figure 3: CRISP-DM

Practical steps to implement CRISP-DM for the airline project

1. Raw Data Hygiene
2. Data Quality & Remediation
3. Bi-variant Associations
4. Data Associations
5. Data Treatments
6. Model Input Testing
7. Feature Engineering
8. Model Benchmarking
9. Model Tuning
10. Model Performance Evaluation
11. Actionable Insights
12. Business Recommendations

Tool & Techniques

The following analytical tools, development tools and machine learning techniques were used to conduct the study analysis for the airline satisfaction prediction project:

1. Anaconda/JupyterLab python integrated development environment (IDE)
2. Pandas and Numpy python libraries are used to organize and manipulate data across the machine learning lifecycle
3. Matplotlib and Seaborn libraries are used to create visualizations of all raw data, preprocessed data and parametric classification results
4. Scikit Learn, XGBoost and scipy stats machine learning libraries are used to build, test and instrument performance results for target variable classification and feature importance (drivers of satisfaction) based on ensemble learning models
5. Ensemble classification algorithms used for model benchmarking, tuning and stacking include, Decision Tree, Bagging, Random Forest, Ada Boost, Gradient Boost and XGBoost
6. GridSearchCV and RandomizedSearchCV are used to perform hyper-parameter tuning of ensemble models

Exploratory Data Analysis

Raw Data & Hygiene

Data Dictionary (Data Source #1)

This data source has several data quality problems which do not warrant direct remediation as the data source is not programmatically used in the analysis. Observations are, however made to extract meaning from this data source by understanding the ordinal rankings (scores) which are used in the subsequent customer survey data which was collected. From the data dictionary, we establish a coherent and standardized 1-5 scoring gradient for individual survey questions to understand the effects on the target variable, satisfaction.

The standardized (generally applied across all sentiment variables) ordinal ranks used for classification modeling are the following:

0. Very Poor
1. Poor
2. Needs Improvement
3. Acceptable
4. Good
5. Excellent

Flight Data (Data Source #2)

This data source provides customer attribute information (e.g., age, gender, loyalty information) in addition to flight information consisting of variables that describe instances of flights taken by the respective customer. Continuous variables include flight distance, departure delay and arrival delay with categorical variables include business class travel and more specific class of service during the flight.

Survey Data (Data Source #3)

This data captures passenger sentiment toward both overall passenger satisfaction as well as individual satisfaction levels with diverse types of airline services e.g., infight service, infight entertainment, seat comfort, app boarding, online support, ease of online booking etc.

Null values are noted in the 3 variables of loyalty, travel type and arrival delay. These null values are replaced with imputed values prior to building and testing a classification model.

Data Quality, Variation & Remediation

Through exploratory data analysis, we make visual observations to gain an intuitive sense of the data prior to modeling. To better understand passenger demographics, below we plot the univariate distribution of customer age separated by satisfaction and nonsatisfaction.

Age Distribution by Satisfaction

From this visual, we are able to determine that the majority of satisfied customers are between 40 and 60 years ago. This information can be used to inform marketing efforts.

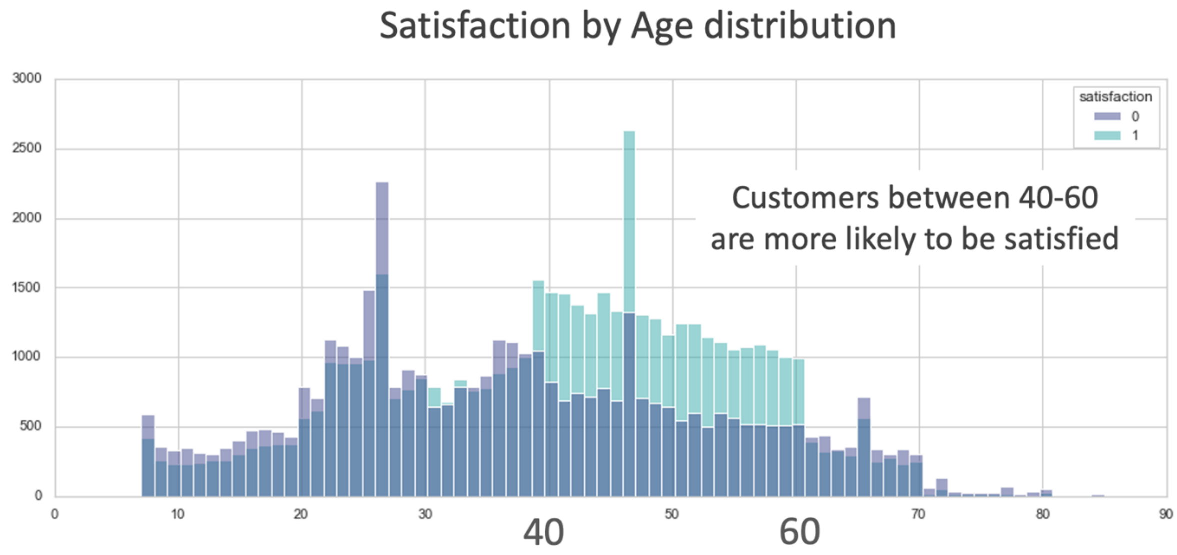


Figure 4: Age distribution by satisfaction

Gender and Customer Loyalty Variables (samples)

Basic demographic information shows an approximately even split between genders, a 3/4 majority of loyalty program customers vs. customers who do not participate in the loyalty program.

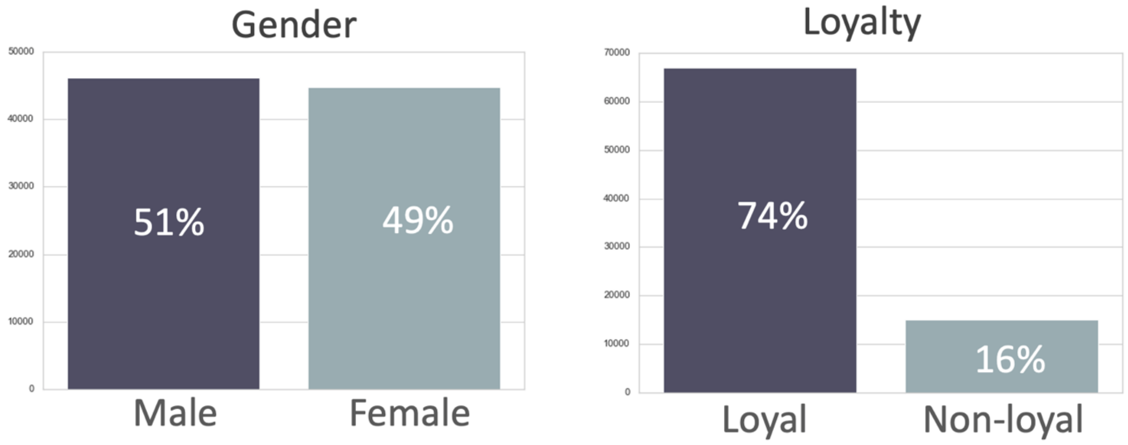


Figure 5: Passenger demographic variable count distributions

Travel Type and Service Class (samples)

A 62% majority of passengers utilize travel for business with specific class of service of the airline ticket purchased accounting for 45% in Economy, 48% in Business and the remaining 7% is Economy+.

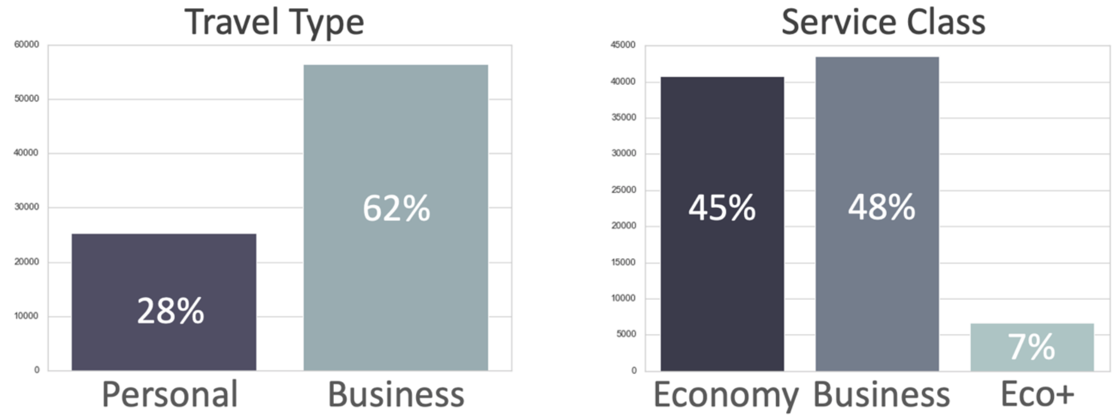


Figure 6: Passenger flight attribute associations

Flight Distance Variable (sample)

Flight distance data is observed to have a large degree of variation and skewness in both departure delay and arrival delay with both containing extreme outliers (a proportionally small number of extremely high values relative to the data set's overall size). The 75th percentile for flight distance is between 2000 and 3000 miles.

This can be viewed in the chart of flight distance below.

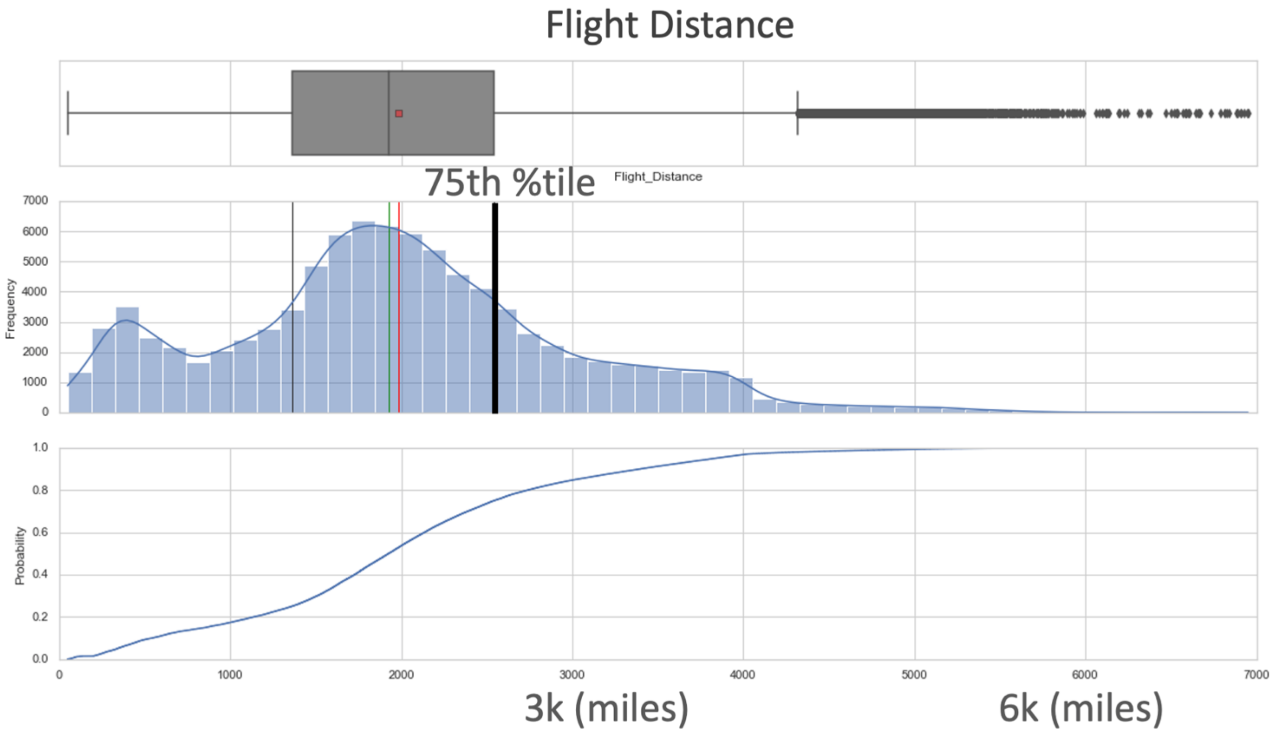


Figure 7: Flight distance distribution and cumulative probability

SURVEY VARIABLES (SAMPLE)

Below are show representative samples of survey variables included in the raw data which describe the customer's sentiment toward two prominent features that influence overall satisfaction, Seat Comfort and In-flight Entertainment. The complete listing of survey variables and their response distributions can be found in the Jupyter notebook which accompanies this report.

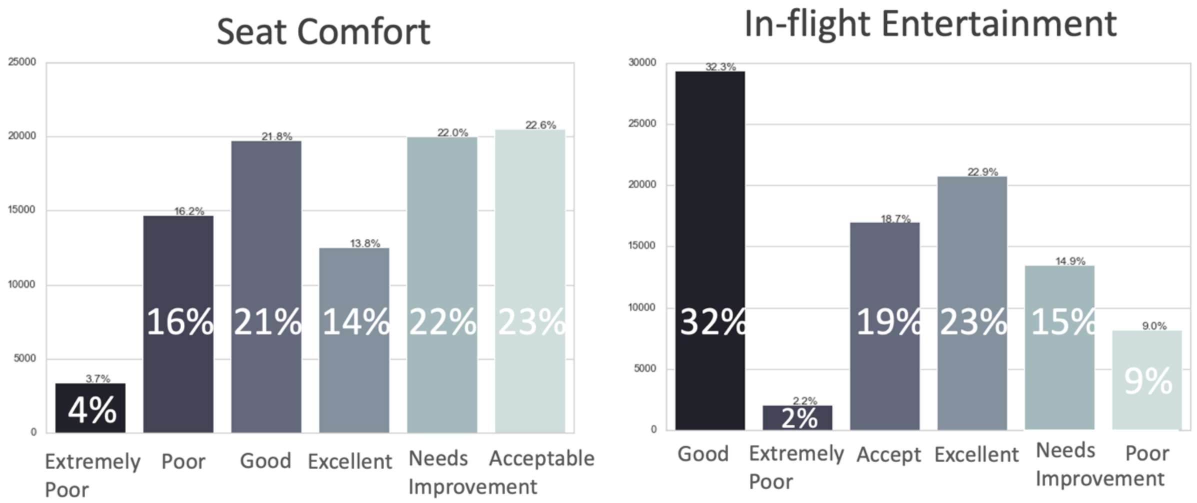


Figure 8: Survey variable samples for seat comfort & in-flight entertainment

BIVARIATE DATA ASSOCIATIONS

Strong collinearity is observed between departure delay and arrival delay for the flight in question. This makes sense given that arrival delay is a function the departure delay. This indicates that one variable should be removed from the data set prior to building a classification model to ensure non-target variables are as independent as possible such that they do not exert an artificially amplified effect on the predicting the target variable.

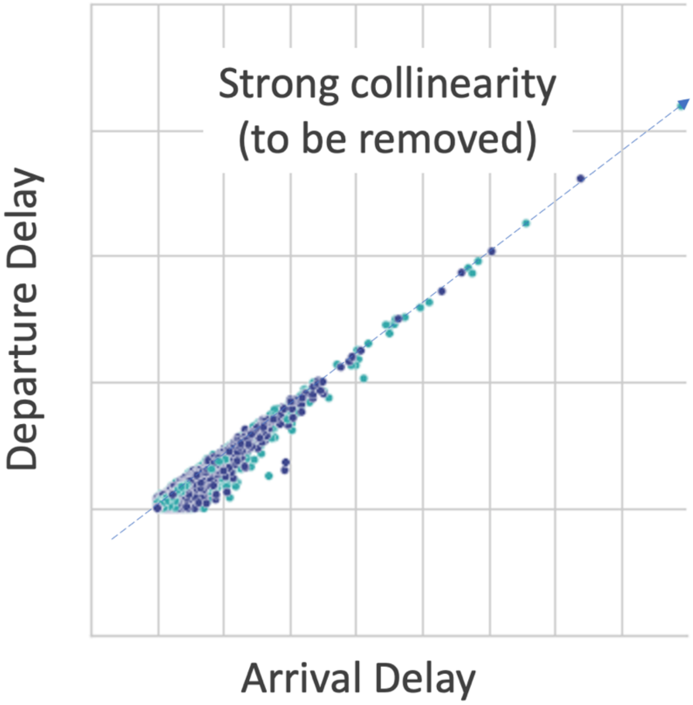
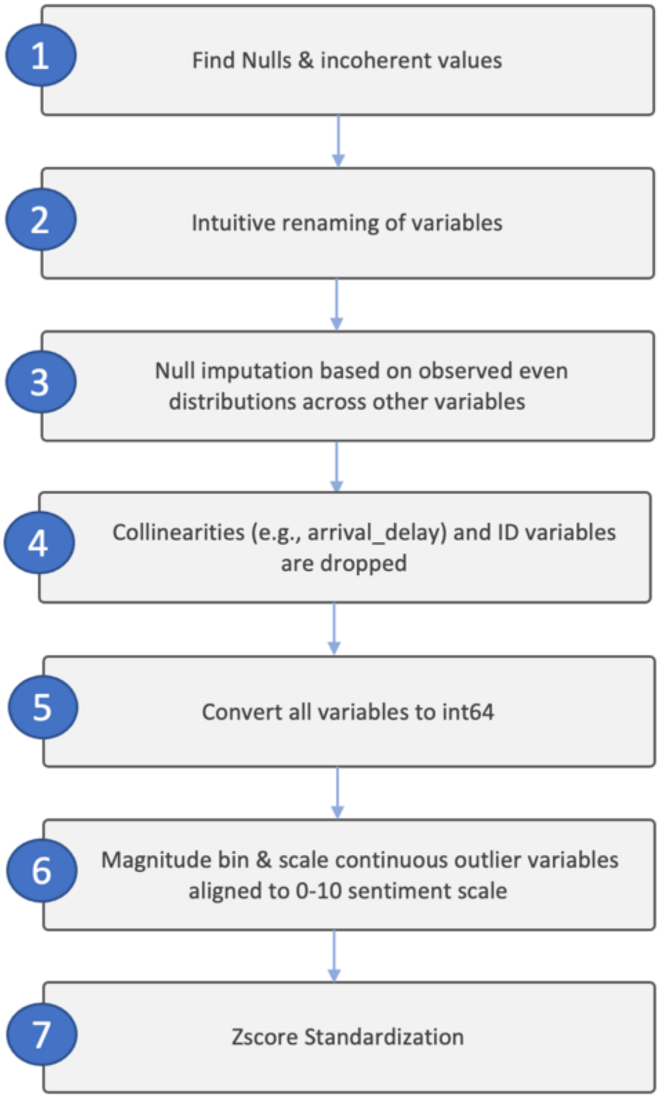


Figure 9: Departure / Arrival Delay Collinearity

Data Description & Preparation

The contextual procedure for data cleaning and preparation for modeling is shown below with the key steps listed in the figure. While the diagram depicts a linear process of sequential steps, in practice the steps are highly iterative with arrows going in both directions. Building programming methods for making this experimentation easy to run and highly visual enables testing a greater (more diverse) number of inputs, models and parameter combinations in the hunt for every possible classification percentage point.

Baseline Data Preprocessing



1 Find Nulls & incoherent values

2 Intuitive renaming of variables

3 Null imputation based on observed even distributions across other variables

4 Collinearities (e.g., arrival_delay) and ID variables are dropped

5 Convert all variables to int64

6 Magnitude bin & scale continuous outlier variables aligned to 0-10 sentiment scale

7 Zscore Standardization

Figure 10: Data cleaning & preparation for modeling

Scaling & Outlier Treatments

The majority of variables obtained consisted of survey features in the range from 0-5 on an ordinal scale. The most pronounced outliers occur in the FLIGHT_DISTANCE, DEPARTURE_DELAY, and ARRIVAL_DELAY variables. Due to collinearity between the two delay variables, ARRIVAL_DELAY is removed noting that this also eliminates the associated null values observed in this variable.

Outliers are removed from the variables using magnitude binning. FLIGHT_DISTANCE and DELAY are magnitude binned with a 5-point range to compress these datasets down into a similar scale and granularity as the 0-5 range survey scores. The approach taken here was to think about possible modes in the data as an order of (negative) magnitude. While it should be acknowledged that we lose information by doing this, we make the independent variable to target variable mapping permutations more discrete.

Flight Delay Magnitude-binned (with collinearity removed)

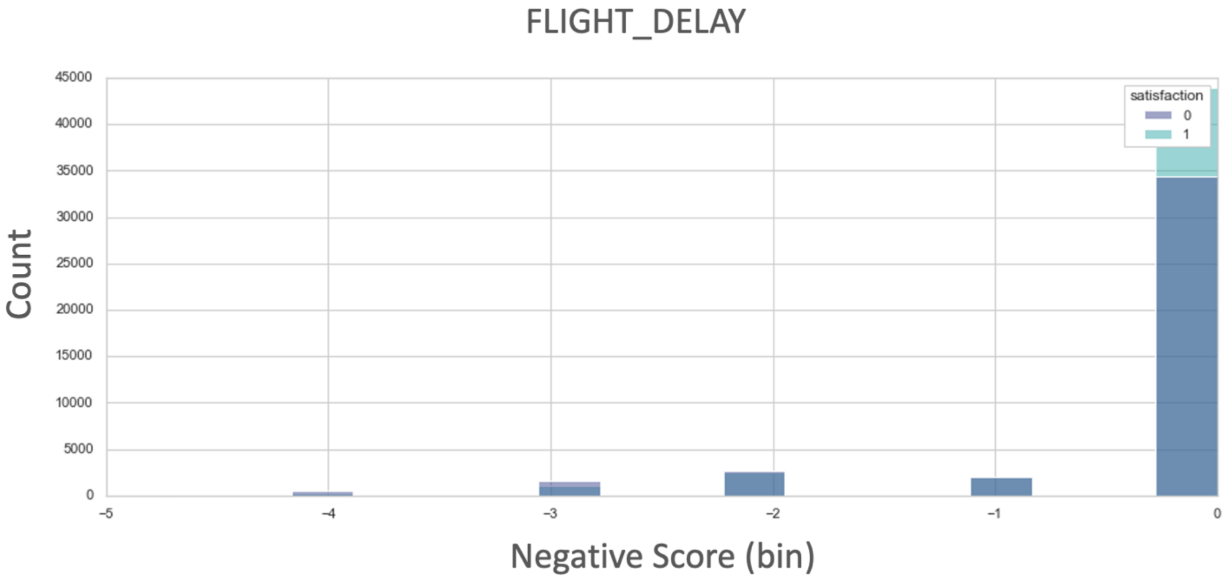


Figure 11: Magnitude binning of flight delay

Survey Score Influence on Satisfaction

Gaining insights from EDA, we observe that overall passenger satisfaction is more likely when passengers also scored INFLIGHT_WIFI_SERVICE and INFLIGHT_ENTERTAINMENT highly on the scale. This provides an intuition that these features potentially have a larger impact on the ability to predict if a customer is satisfied with their airline experience.

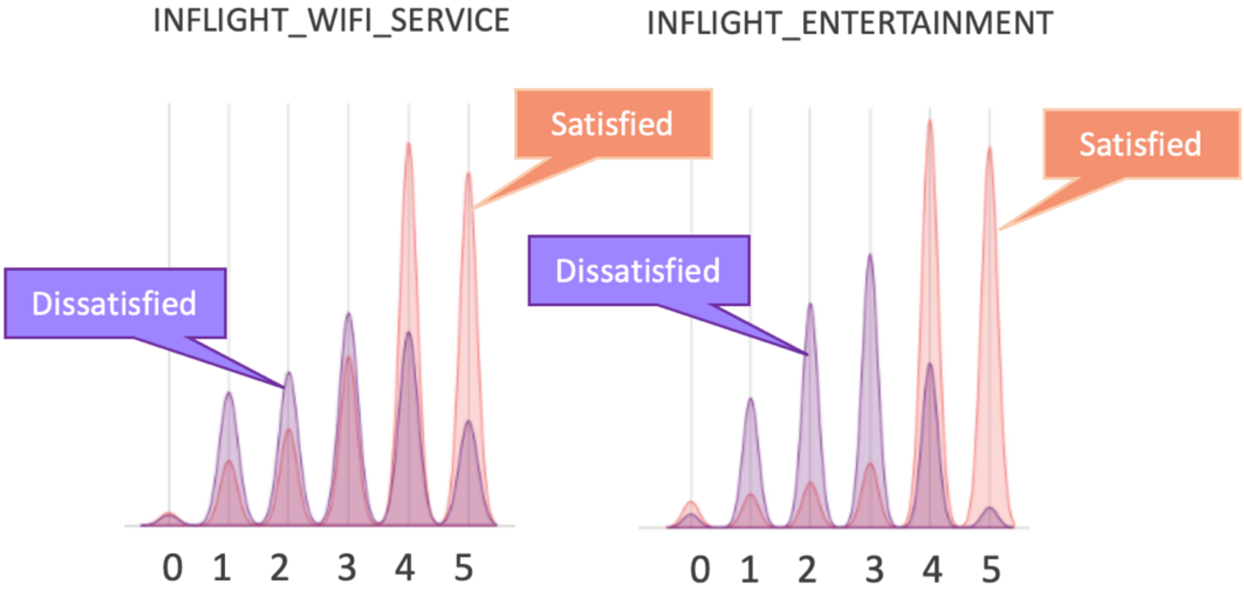


Figure 12: Sentiment variable associations with target variable

Preprocessed Features

The following preprocessed features are used for benchmarking and final classification modeling. We note the target variable as SATISFACTION and list three distinct types of data about the customer. Customer attributes describe the customers themselves and have values which can change slowly over time (e.g., AGE) or may never change (e.g., FEMALE). Features are present which describe the actual flight taken by the customer in addition to the sentiment variables which capture the customer's satisfaction ratings for individual aspects of their customer experience.

Target / Prediction Attribute	
SATISFACTION	

Customer Attribute Features (zscores)	
FEMALE	LOYALTY_SCORE
AGE	

Flight Attribute Features (zscores)	
BUSINESS_TRAVEL	SERVICE_CLASS
FLIGHT_DISTANCE	DEPARTURE_DELAY

Survey Sentiment Features (zscores)	
APP_BOARDING_SCORE	ONLINE_BOOKING_EASE_SCORE
CHECKIN_SERVICE_SCORE	LEG_ROOM_SCORE
INFLIGHT_CLEANLINESS_SCORE	BAGGAGE_HANDLING_SCORE
SEAT_COMFORT_SCORE	FLIGHT_TIME_CONVENIENCE_SCORE
INFLIGHT_WIFI_SERVICE_SCORE	FOOD_AND_DRINK_SCORE
INFLIGHT_ENTERTAINMENT_SCORE	INFLIGHT_SERVICE_SCORE
ONLINE_SUPPORT_SCORE	GATE_LOCATION_SCORE

Table 1: Model Input Features

Model Input Dataset (boxplot)

Model inputs are binned and scaled using ZScore standardization as shown in the below figure. Performing this transformation on the data establishes a new mean of 0 and a standard deviation to 1. All feature value ranges are to a similar scale in aiming to make the classification model more effective in predicting satisfaction.

The boxplot below shows the central tendency and spread for all input features on a -4 to 8 scale. The survey (sentiment) variables range between 0 and 5.

Boxplot of standardized model input features

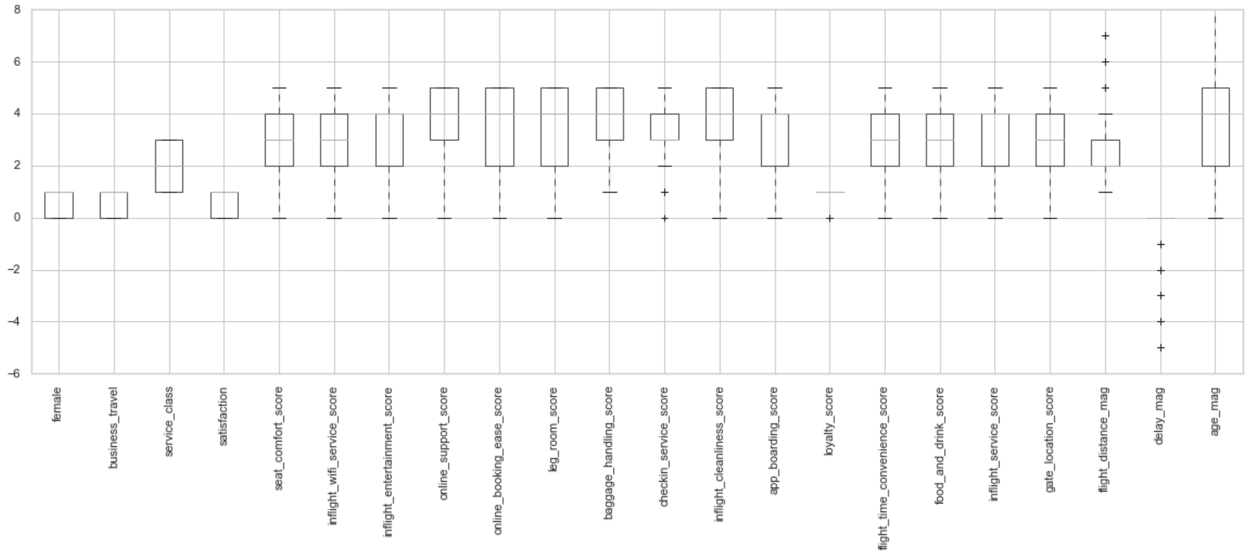
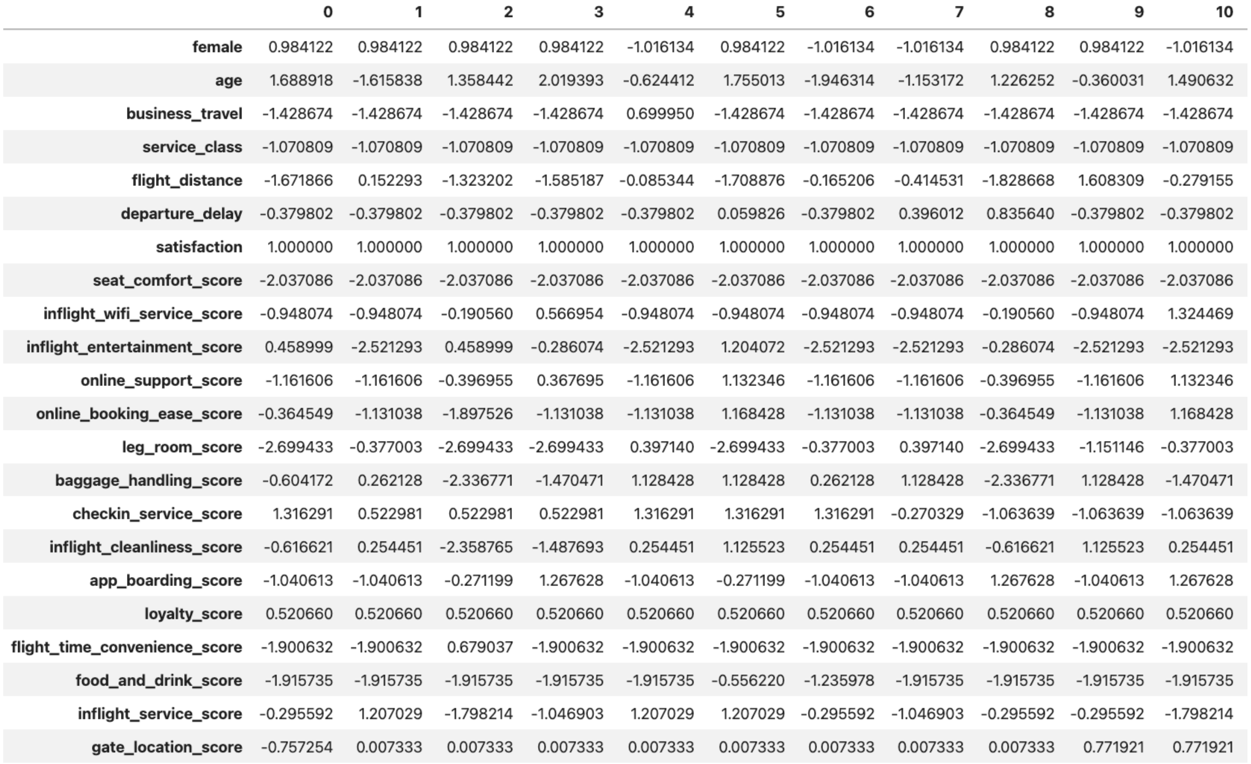


Figure 13: Boxplot of zscore standardized input data selected for modeling

Model Input Data

Below we observe the floating-point numbers generated through ZScore standardization and which serve as the 21 independent features and basis for classification model fitting. Note that we have converted the raw data into a pure floating-point matrix of numbers crafted explicitly to map to the binary target variable, Satisfaction.

ZScore Standardized Input Dataset (sample)



	0	1	2	3	4	5	6	7	8	9	10
female	0.984122	0.984122	0.984122	0.984122	-1.016134	0.984122	-1.016134	-1.016134	0.984122	0.984122	-1.016134
age	1.688918	-1.615838	1.358442	2.019393	-0.624412	1.755013	-1.946314	-1.153172	1.226252	-0.360031	1.490632
business_travel	-1.428674	-1.428674	-1.428674	-1.428674	0.699950	-1.428674	-1.428674	-1.428674	-1.428674	-1.428674	-1.428674
service_class	-1.070809	-1.070809	-1.070809	-1.070809	-1.070809	-1.070809	-1.070809	-1.070809	-1.070809	-1.070809	-1.070809
flight_distance	-1.671866	0.152293	-1.323202	-1.585187	-0.085344	-1.708876	-0.165206	-0.414531	-1.828668	1.608309	-0.279155
departure_delay	-0.379802	-0.379802	-0.379802	-0.379802	-0.379802	0.059826	-0.379802	0.396012	0.835640	-0.379802	-0.379802
satisfaction	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
seat_comfort_score	-2.037086	-2.037086	-2.037086	-2.037086	-2.037086	-2.037086	-2.037086	-2.037086	-2.037086	-2.037086	-2.037086
inflight_wifi_service_score	-0.948074	-0.948074	-0.190560	0.566954	-0.948074	-0.948074	-0.948074	-0.948074	-0.190560	-0.948074	1.324469
inflight_entertainment_score	0.458999	-2.521293	0.458999	-0.286074	-2.521293	1.204072	-2.521293	-2.521293	-0.286074	-2.521293	-2.521293
online_support_score	-1.161606	-1.161606	-0.396955	0.367695	-1.161606	1.132346	-1.161606	-1.161606	-0.396955	-1.161606	1.132346
online_booking_ease_score	-0.364549	-1.131038	-1.897526	-1.131038	-1.131038	1.168428	-1.131038	-1.131038	-0.364549	-1.131038	1.168428
leg_room_score	-2.699433	-0.377003	-2.699433	-2.699433	0.397140	-2.699433	-0.377003	0.397140	-2.699433	-1.151146	-0.377003
baggage_handling_score	-0.604172	0.262128	-2.336771	-1.470471	1.128428	1.128428	0.262128	1.128428	-2.336771	1.128428	-1.470471
checkin_service_score	1.316291	0.522981	0.522981	0.522981	1.316291	1.316291	1.316291	-0.270329	-1.063639	-1.063639	-1.063639
inflight_cleanliness_score	-0.616621	0.254451	-2.358765	-1.487693	0.254451	1.125523	0.254451	0.254451	-0.616621	1.125523	0.254451
app_boarding_score	-1.040613	-1.040613	-0.271199	1.267628	-1.040613	-0.271199	-1.040613	-1.040613	1.267628	-1.040613	1.267628
loyalty_score	0.520660	0.520660	0.520660	0.520660	0.520660	0.520660	0.520660	0.520660	0.520660	0.520660	0.520660
flight_time_convenience_score	-1.900632	-1.900632	0.679037	-1.900632	-1.900632	-1.900632	-1.900632	-1.900632	-1.900632	-1.900632	-1.900632
food_and_drink_score	-1.915735	-1.915735	-1.915735	-1.915735	-1.915735	-0.556220	-1.235978	-1.915735	-1.915735	-1.915735	-1.915735
inflight_service_score	-0.295592	1.207029	-1.798214	-1.046903	1.207029	1.207029	-0.295592	-1.046903	-0.295592	-0.295592	-1.798214
gate_location_score	-0.757254	0.007333	0.007333	0.007333	0.007333	0.007333	0.007333	0.007333	0.007333	0.771921	0.771921

Table 2: Transposed standardized input data set for modeling (sample of first 10 rows)

Classification Modeling

Competing Ensemble Classifiers

Given the experimental nature of observing how different classification models perform on Train and Test input data, various combinations of model sequences were run and filtered down to the best 3 models based on the metric we want to maximize and tighten, Precision. In a sense, the models built and run compete against each other to see which among them will be the strongest learner(s).

Individual Benchmark Model Sequence

In the initial benchmarking sequence, we test 4 different preprocessed variants of model input data and execute them individually against the following baseline classifiers: Decision Tree, Bagging, Random Forest, Ada Boost, Gradient Boost and XGBoost. The benchmark sequence of individual models is shown in the diagram below.

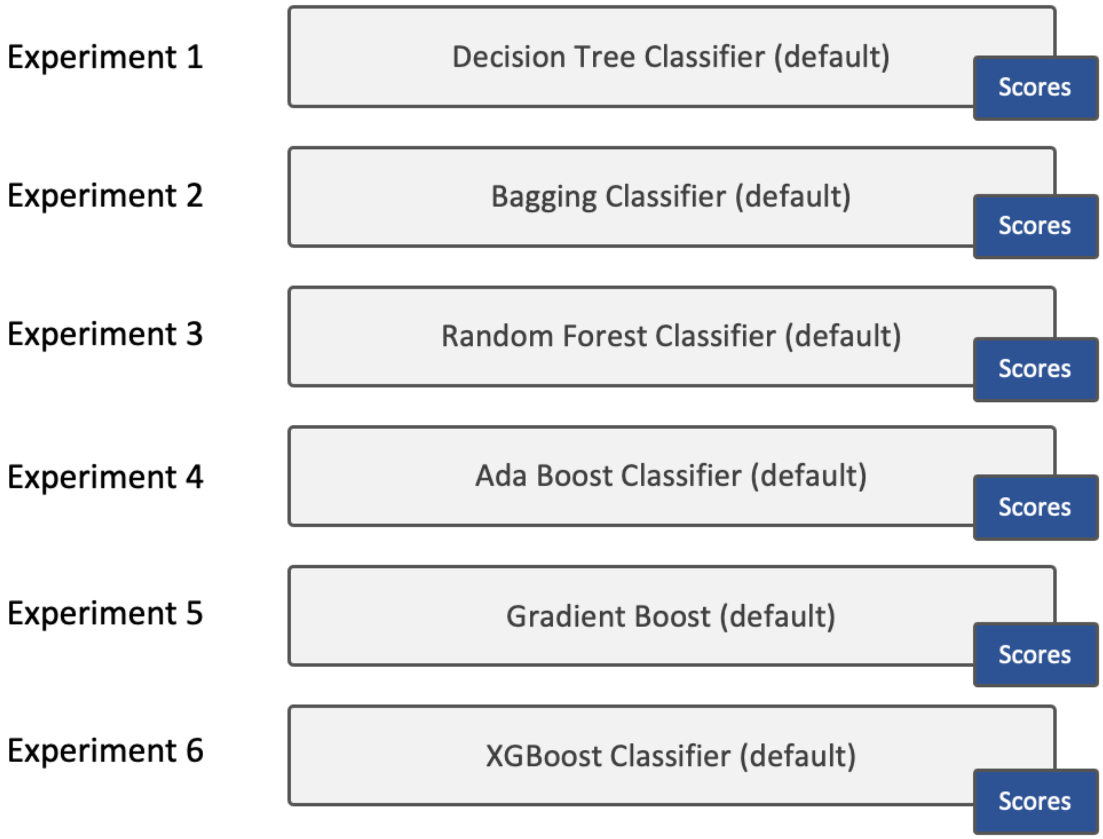

Experiment 1
Decision Tree Classifier (default)
Scores
Experiment 2
Bagging Classifier (default)
Scores
Experiment 3
Random Forest Classifier (default)
Scores
Experiment 4
Ada Boost Classifier (default)
Scores
Experiment 5
Gradient Boost (default)
Scores
Experiment 6
XGBoost Classifier (default)
Scores

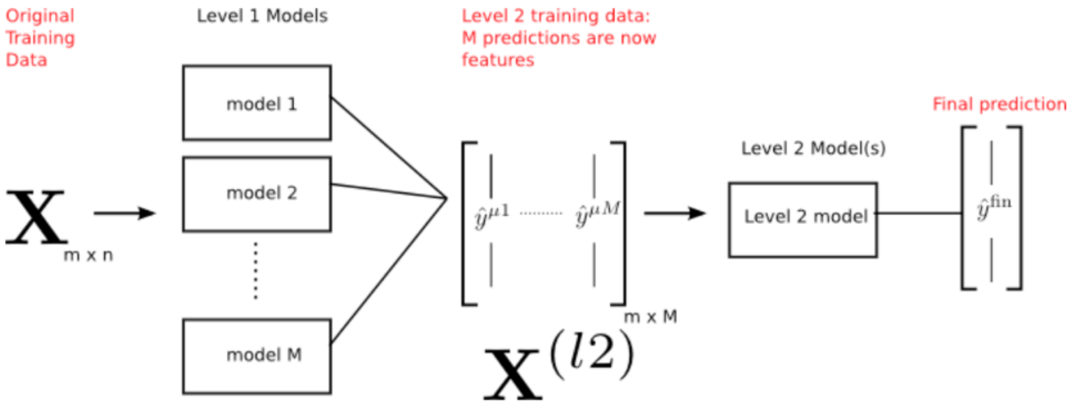
Figure 14: Baseline benchmark sequence of models fit & scored on Train/Test data

From the initial model sequence, it is observed that the 3 best individual models listed are: Random Forest (Precision T/T = 1.0/.96), Bagging (Precision T/T = 1.0/.96) and XGBoost (Precision T/T = .98/.96). These models are selected based on their ability to maximize the out-of-sample Precision score while keeping the delta between in-sample and out-of-sample data as small as possible. We observe that individual ensemble models with default settings perform well without hyperparameter tuning.

Ensemble Classifier Stacking

We employ stacking of ensemble models for the purpose of achieving robustness in classification performance scoring. Individual binary classification models possess different strengths and weaknesses give the data set provided. The overall objective of stacking is for different models to supplement each other in detecting patterns and associations in cases where peer models may be less capable and effective in doing so.

The diagram below illustrates the general concept of how stacking is implemented:



Source: www.kdnuggets.com/2017/02/stacking-models-imropved-predictions.html

Figure 15: Model stacking algorithmic concept

From among the 3 strongest learners high-lighted in the benchmarking sequence, we then stack these models and rotate each of them to serve as the final meta model in the stacked sequence.

Benchmark Stacking on Best 3 Base Models

Below shows the ensemble stacking assembly configured from the best 3 individual models identified in the prior section. While the three best individual models are selected for the level 1 of stacking, the same 3 individual models are rotated as the final (meta) model to characterize the classification performance differences in among executing them in the different combinations. The objective in doing this is to find the optimal sequence assembly which achieves the highest Precision score on Test data with the smallest delta between results from Train and Test data.

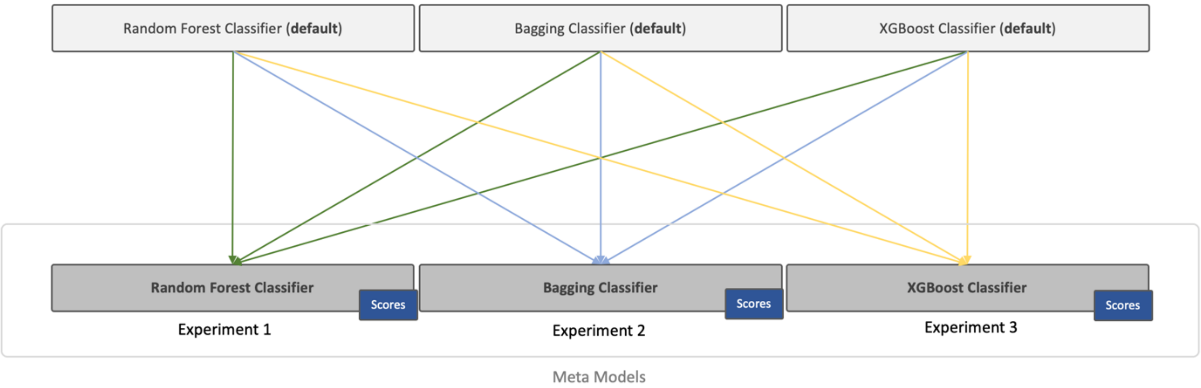


Figure 16: Benchmark model stacking configuration

Model Performance Evaluation

A key aspect of the modeling approach is to determine which classification measure to maximize the value of through tuning and experimentation to select final model(s) from which to understand customer sentiment drivers.

The following measures are considered to evaluate model performance:

- **Accuracy** = (True Positives + True Negatives) / Total
- **Recall** = True Positives / (True Positives + False Negatives)
- **Precision** = True Positives / (True Positives + False Positives)

Recall should be used when False Negatives are very expensive e.g., loan default. Precision should be used when False Positives are very expensive e.g., drone strike or losing unhappy customers. Maximizing precision will minimize the number false positives, whereas maximizing the recall will minimize the number of false negatives.

Now let us ask the question of which performance metric is the most appropriate in the context of Falcon Airlines. In doing so, we question which condition is more expensive to the business between False Positives and False Negatives (False Positive corresponds to assuming a customer is satisfied when they are in fact not satisfied).

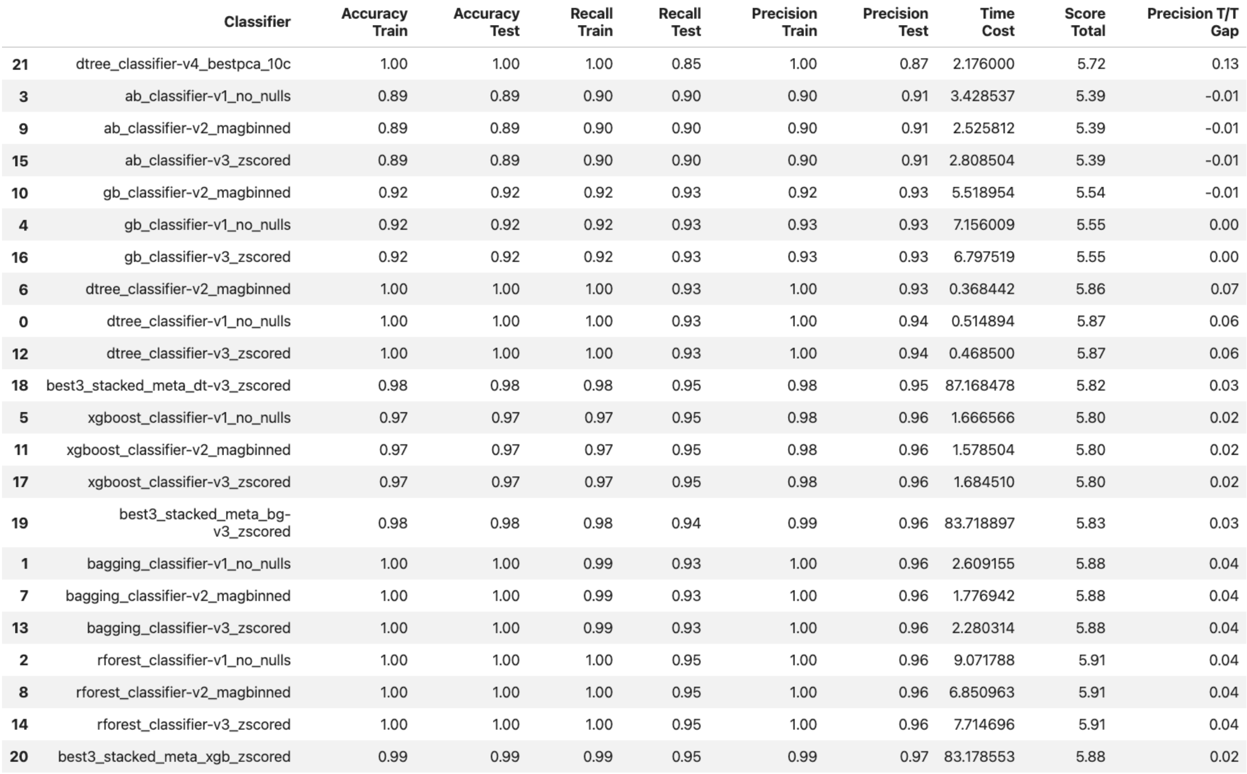
When a satisfied customers is incorrectly classified as an unsatisfied customer, then marketing efforts are ineffective and marketing investments will be wasted with a negative economic impact proportional to a percentage of the budget.

When an unsatisfied customer is incorrectly classified as a satisfied customer, then they are more likely to leave and move to a competitor as marketing and operational investments will not reach them to motivate loyalty through an improved customer experience message.

Risk of attrition (customer moves to a competitor airline) has the largest negative impact to the business and so we thus select **Precision** as the primary model performance metric to optimize for Falcon Airlines.

Furthermore, we must remain mindful of overfitting of the classification model leading to degraded performance in the future against prior unseen data. In order to devise business insights and recommendations, model sequencing and hyper-parameter tuning is performed with the objective of *maximizing Precision score* while simultaneously *keeping the delta between Train and Test data as small as possible* to minimize overfitting.

Benchmark Classification Results



	Classifier	Accuracy Train	Accuracy Test	Recall Train	Recall Test	Precision Train	Precision Test	Time Cost	Score Total	Precision T/T Gap
21	dtree_classifier-v4_bestpca_10c	1.00	1.00	1.00	0.85	1.00	0.87	2.176000	5.72	0.13
3	ab_classifier-v1_no_nulls	0.89	0.89	0.90	0.90	0.90	0.91	3.428537	5.39	-0.01
9	ab_classifier-v2_magbinned	0.89	0.89	0.90	0.90	0.90	0.91	2.525812	5.39	-0.01
15	ab_classifier-v3_zscored	0.89	0.89	0.90	0.90	0.90	0.91	2.808504	5.39	-0.01
10	gb_classifier-v2_magbinned	0.92	0.92	0.92	0.93	0.92	0.93	5.518954	5.54	-0.01
4	gb_classifier-v1_no_nulls	0.92	0.92	0.92	0.93	0.93	0.93	7.156009	5.55	0.00
16	gb_classifier-v3_zscored	0.92	0.92	0.92	0.93	0.93	0.93	6.797519	5.55	0.00
6	dtree_classifier-v2_magbinned	1.00	1.00	1.00	0.93	1.00	0.93	0.368442	5.86	0.07
0	dtree_classifier-v1_no_nulls	1.00	1.00	1.00	0.93	1.00	0.94	0.514894	5.87	0.06
12	dtree_classifier-v3_zscored	1.00	1.00	1.00	0.93	1.00	0.94	0.468500	5.87	0.06
18	best3_stacked_meta_dt-v3_zscored	0.98	0.98	0.98	0.95	0.98	0.95	87.168478	5.82	0.03
5	xgboost_classifier-v1_no_nulls	0.97	0.97	0.97	0.95	0.98	0.96	1.666566	5.80	0.02
11	xgboost_classifier-v2_magbinned	0.97	0.97	0.97	0.95	0.98	0.96	1.578504	5.80	0.02
17	xgboost_classifier-v3_zscored	0.97	0.97	0.97	0.95	0.98	0.96	1.684510	5.80	0.02
19	best3_stacked_meta_bg-v3_zscored	0.98	0.98	0.98	0.94	0.99	0.96	83.718897	5.83	0.03
1	bagging_classifier-v1_no_nulls	1.00	1.00	0.99	0.93	1.00	0.96	2.609155	5.88	0.04
7	bagging_classifier-v2_magbinned	1.00	1.00	0.99	0.93	1.00	0.96	1.776942	5.88	0.04
13	bagging_classifier-v3_zscored	1.00	1.00	0.99	0.93	1.00	0.96	2.280314	5.88	0.04
2	rforest_classifier-v1_no_nulls	1.00	1.00	1.00	0.95	1.00	0.96	9.071788	5.91	0.04
8	rforest_classifier-v2_magbinned	1.00	1.00	1.00	0.95	1.00	0.96	6.850963	5.91	0.04
14	rforest_classifier-v3_zscored	1.00	1.00	1.00	0.95	1.00	0.96	7.714696	5.91	0.04
20	best3_stacked_meta_xgb_zscored	0.99	0.99	0.99	0.95	0.99	0.97	83.178553	5.88	0.02

Table 3: Preprocessing Variant/model combination scoring

Benchmarking and Stacking Classification Results

Below we can observe Accuracy, Recall and Precision scores for 22 different experiments. 4 different variants of preprocessed input data were fit on 7 different baseline ensemble models with 1 additional test performed using Principal Component Analysis (PCA) features as inputs (in the last slot). We can observe that using the PCA inputs produced the lowest Precision scores on Train and Test data and therefore was not pursued further in the model tuning and selection process. The gray boxes on the chart indicate the benchmark models which produced the highest Precision scores on Test data combined with the smallest delta on Precision on Train and Test data.

These 3 model/input variant combinations with the best Precision score in benchmarking include:

- Random Forest classifier with zscore standardized inputs
- Bagging classifier with zscore standardized inputs
- XGBoost classifier with zscore standardized inputs

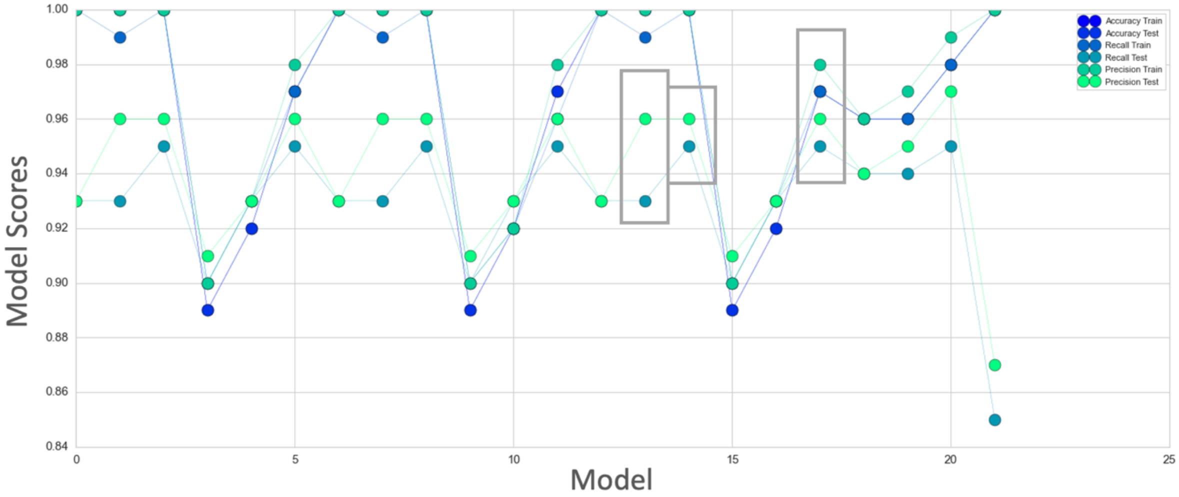


Figure 17: Benchmark & stacking ensemble classification model scores

Benchmark Classification Runtime Cost Results

The runtime costs for the benchmarking sequences were all under 40 seconds in length operating on over 90,000 rows of input data. This initial benchmark of time for each model or sequence to fit, predict and score shows that running any of them in production is feasible from a speed of execution standpoint. The 3 best individual classifiers are shown in the gray boxes below (Random Forest, Bagging and XGBoost). The amount of time taken for the stacking sequences are shown as the highest 3 bars at model index 18, 19, and 20 below.

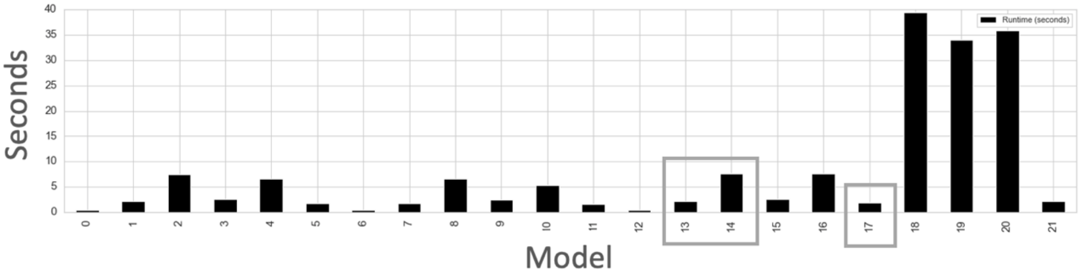
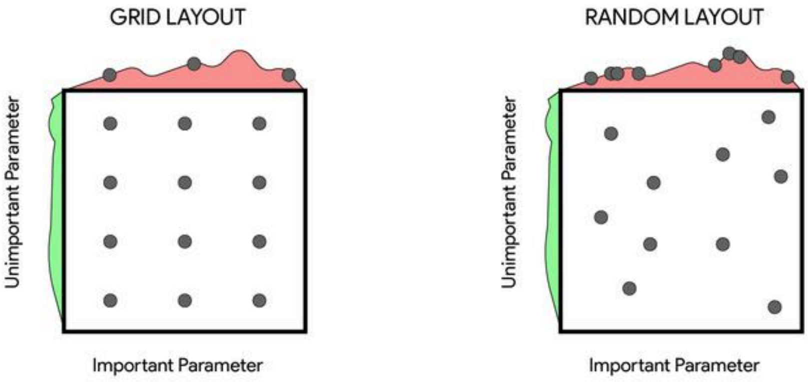


Figure 18: Benchmark & stacking ensemble classification model runtimes

Classification Tuning

GridSearchCV and RandomizedSearchCV algorithms are employed to experiment with hyper-parameter tuning with an objective to find a model and combination of hyper-parameters that improve upon the Precision scores obtained in the benchmarking individual and stacked sequence results.



Source: www.geeksforgeeks.org/hyperparameters-optimization-methods-ml

Figure 19: Grid search and randomized search

It should be noted that there is a high time cost associated with these hyper-parameter tuning methods due to the potentially large number of parameter combinations used when searching for the best model configuration.

Tuned Classifier Stacking

Below shows the ensemble stacking assembly configured after the best 3 individual models that were tuned using both GridSearchCV and RandomizedSearchCV algorithms described above. From these experiments we can observe out-of-sample Precision scores for 3 different stacking trials noted in the diagram below.

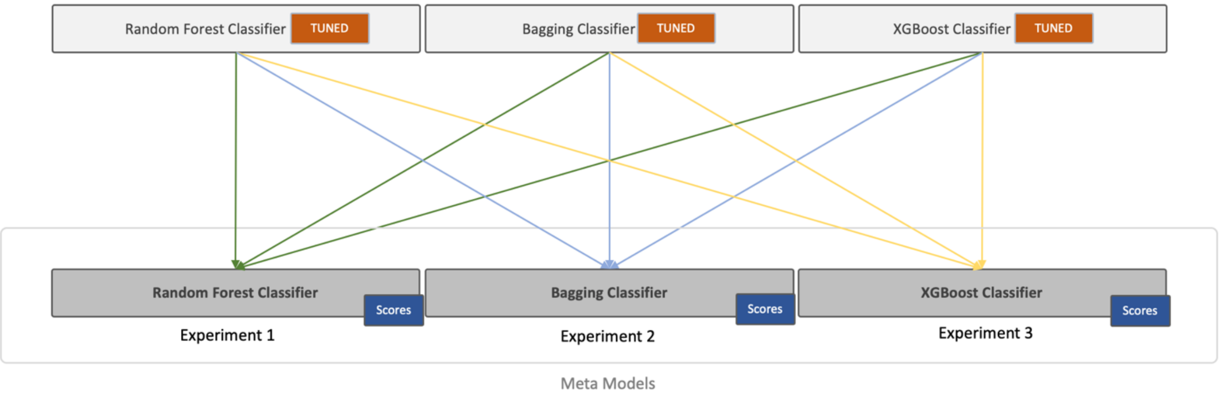
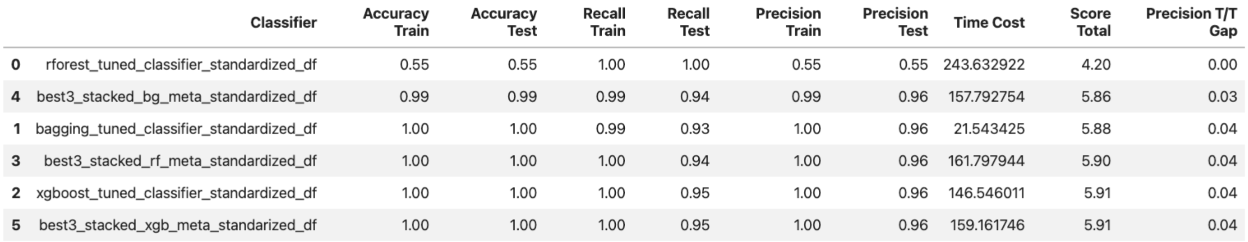


Figure 20: Tuned models stacked with each of best 3 level 2 models

Stacking results on best 3 tuned models

The below table shows the classification performance results for individual tuned classifiers (Random Forest, Bagging and XGBoost) as well as the stacked assemblies rotating each of the best 3 individual models as the level 2 meta classifier. It can be observed in the results below that both the XGBoost tuned classifier and also the best 3 model stacked assembly with XGBoost level 2 classifier produces the best Precision score of 1.0 (in-sample) and .96 (out-of-sample). While these scores reasonably good, the delta between Train and Test is larger than the same stacked assembly using a default level 1 and level 2 models.



	Classifier	Accuracy Train	Accuracy Test	Recall Train	Recall Test	Precision Train	Precision Test	Time Cost	Score Total	Precision T/T Gap
0	rforest_tuned_classifier_standardized_df	0.55	0.55	1.00	1.00	0.55	0.55	243.632922	4.20	0.00
4	best3_stacked_bg_meta_standardized_df	0.99	0.99	0.99	0.94	0.99	0.96	157.792754	5.86	0.03
1	bagging_tuned_classifier_standardized_df	1.00	1.00	0.99	0.93	1.00	0.96	21.543425	5.88	0.04
3	best3_stacked_rf_meta_standardized_df	1.00	1.00	1.00	0.94	1.00	0.96	161.797944	5.90	0.04
2	xgboost_tuned_classifier_standardized_df	1.00	1.00	1.00	0.95	1.00	0.96	146.546011	5.91	0.04
5	best3_stacked_xgb_meta_standarized_df	1.00	1.00	1.00	0.95	1.00	0.96	159.161746	5.91	0.04

Table: Results for stacking of tuned models

Best Model Assembly

These final-pass results based on hyper-parameter tuning and stacking were not observed to be significantly better than results obtained with a default stacking assembly of Random Forest, Bagging and XGBoost with an XGBoost classifier as the level 2 meta model. This machine learning configuration is capable of producing an in-sample Precision score of .99 with an out-of-sample Precision score of .97. The results able to be obtained from hyper-parameter tuning of the respective models with GridSearchCV and RandomizedSearchCV resulted in Precision scores of 1.0 (in-sample) and .96 (out-of-sample). Given that the delta between Train and Test results becomes wider using the tuned models, we select and recommend the benchmark stacking sequence mentioned directly above for deployment and integration with business operations. In doing so, we expect to misclassify approximately 10 out of every 1000 customers as being satisfied with the airline based on partial data when customer satisfaction scoring is not available. We assert here that this is an acceptable level of risk to recommendation-based future investments and business process improvements.

Actionable Insights

Classification Confusion Matrix

The confusion matrix below shows the actual classification results in terms of counts and percentages of true and false positives and negatives.

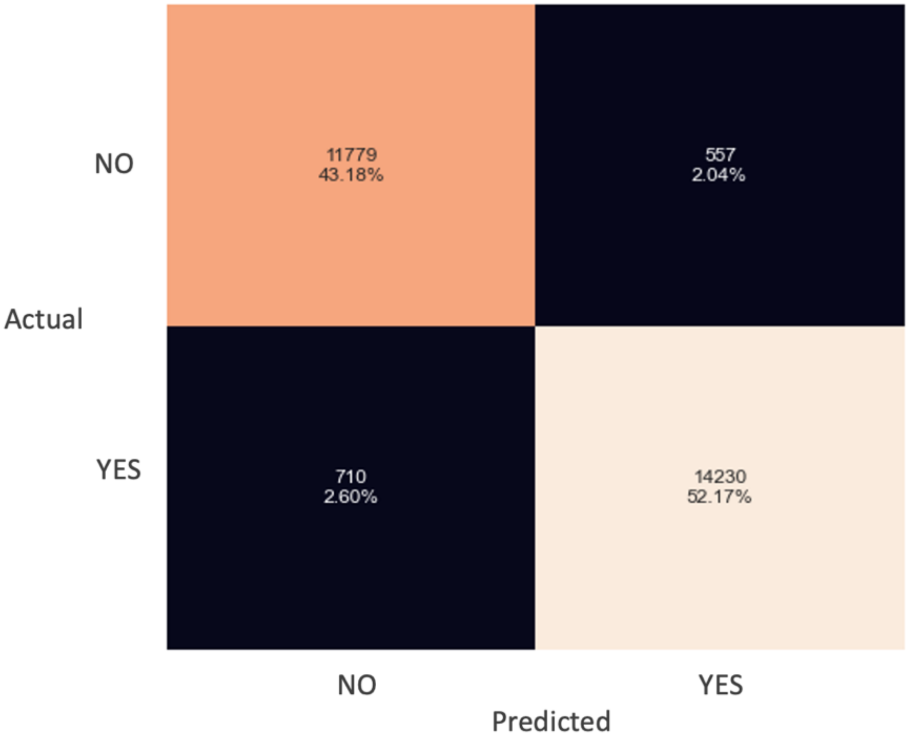


Figure 21: Satisfaction Classification Confusion Matrix

Precision vs. Recall

Precision/Recall Balance is showed in the chart below. While we have optimized for Precision based on the business context, we also desire to achieve a high Recall score to increase robustness and more an improved score balance by reducing false negatives. Reducing the false negatives will aid in maximizing each marketing dollar targeting customer segments. Note: this study can be extended further by performing a clustering / customer segmenting prior to building and testing classification model sequences.

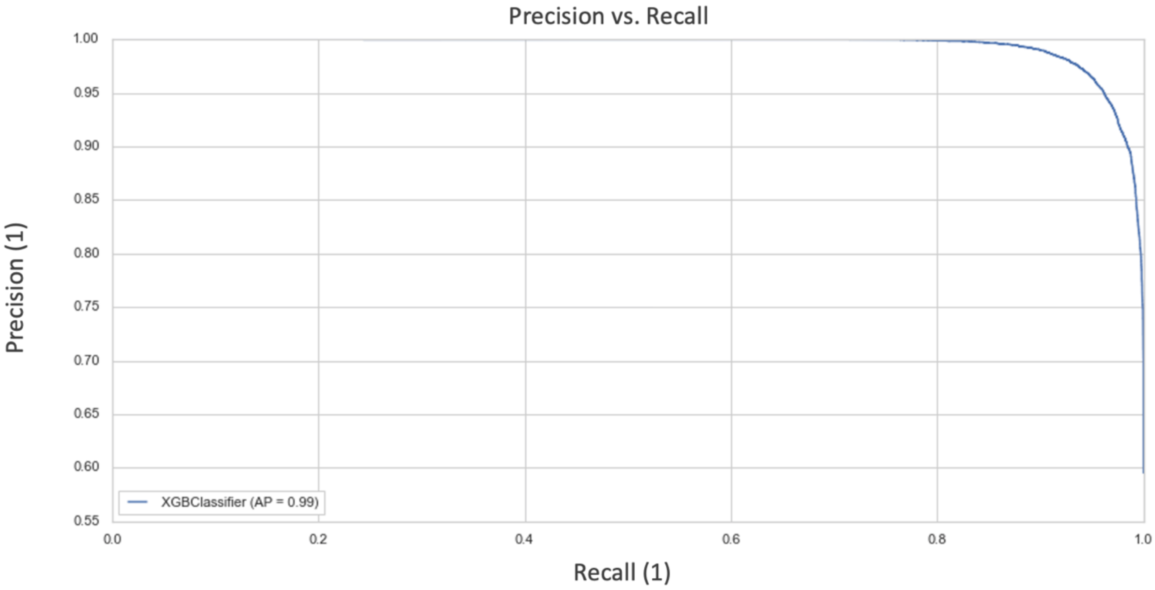


Figure 22: Precision vs. Recall (XGBClassifier)

ROC / AUC Performance

Classification performance can be visualized and understood using a Receiver Operating Characteristic (ROC) and area under the curve (AUC) shown below. We can observe the dotted red line, which represents a random classifier that would be expected to make accurate class predictions 50% of the time on average. The best XGBoost classifier covers 99% of the area under the curve and therefore confirms a strong learner that can be used for production use cases. From what is observed, we can reasonably assert that the survey / sentiment features are a good predictor of overall customer satisfaction and are not strongly influenced by variations in preprocessing data treatments.

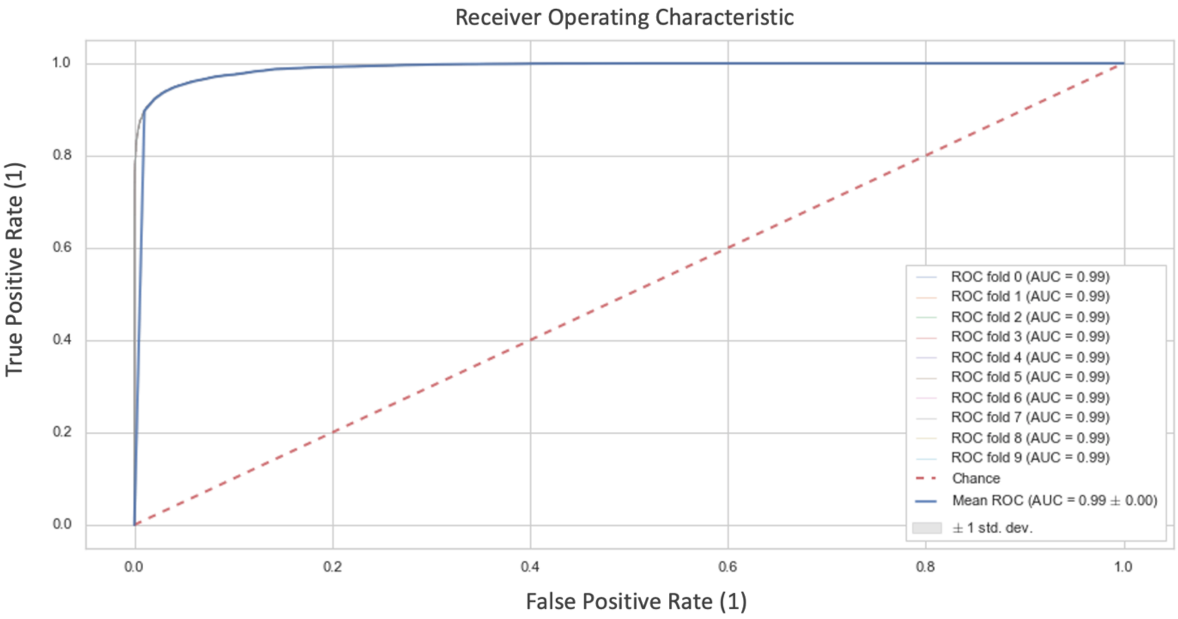


Figure 23: ROC/AUC Classification Performance (XGBoost)

Classification Efficacy

The ensemble methods employed in the analysis were able to achieve an out-of-sample Precision score of .97 and an in-sample Precision score of .99. We conclude that satisfaction classification efficacy is high in the analysis performed. The best 3 stacked models with an XGBoost meta model implement a strong machine learner that is feasible to be automated, scheduled and run in a product environment or on demand as new survey data is collected.

While a production-grade leaner has been implemented, further optimizations are possible to increase Precision and/or other performance metrics with more variations of ML pipeline experiments.

Satisfaction Feature Importance

Below we can observe that the top 10 features which contribute most to the binary classification. These features highlight tangible areas within the business that are actionable and are likely to affect customer satisfaction at present and in the near future. Falcon Airlines should leverage and integrate this information into its business strategy to target objectives across functional business units. The chart below shows these top 10 feature importances.

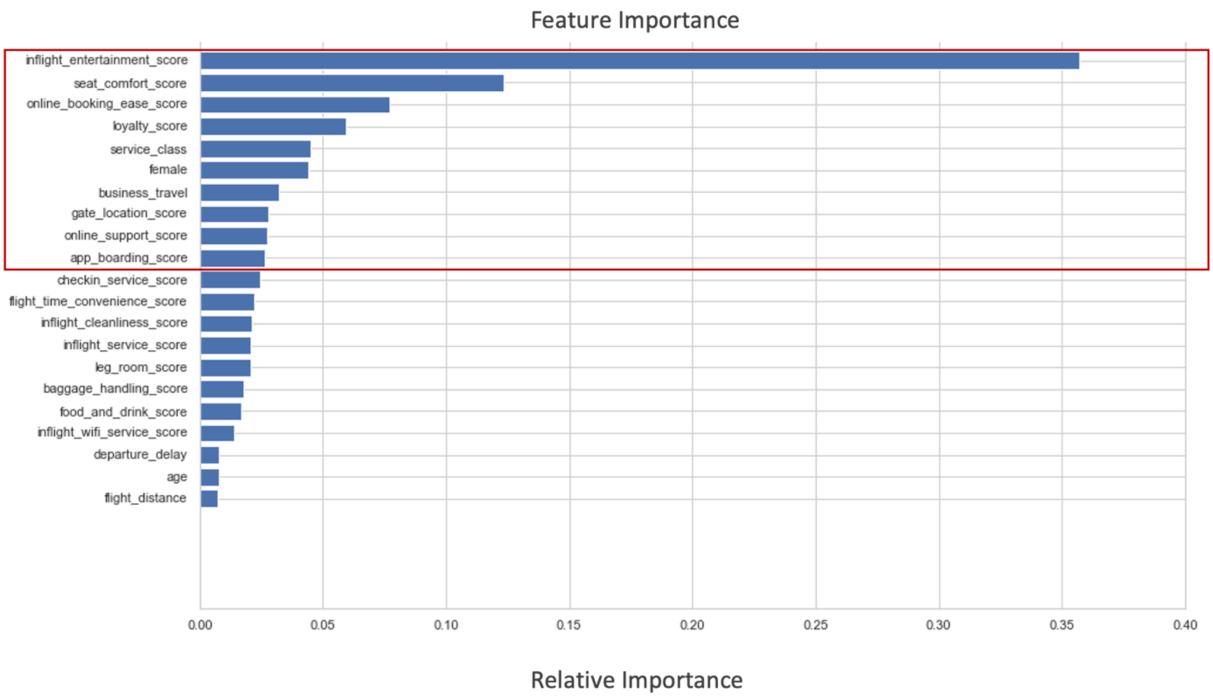


Figure 24: Relative feature importances

Business Recommendations

The business recommendations below present the realization of the study and this should be applied to strategy and change management for Falcon Airlines. Each recommendation will require a budget, resource allocation and measurement system to monitoring return on investments. Business unit leaders are intended to drive the following improvements.

Below is provided a breakdown of specific improvement or investment recommendations which address the top 10 list above.

Rank	Strategy Area	Component	Recommendation
1	Comfort & Entertainment Services	Inflight Entertainment	Audit aircraft for outdated entertainment equipment, upgrade all outdated equipment
2	Aircraft Design & Accommodations	Seat Comfort	Audit aircraft for damaged or outdated seats, upgrade/replace or upgrade seat comfort
3	IT Systems	Online Booking Experience	Performance user experience analysis for online booking flow, scope project to upgrade or replace user experience, model customer journey
4	Marketing, Customer Programs	Loyalty Program	Run focus group, solicit customer feedback, run design thinking workshop to disrupt loyalty programs as we know them. Design and run disruptive story-telling marketing campaign
5	Marketing, Customer Programs	Service Class	Review classes of service design and experiment with new perks/offerings to maximize per class margins. Run story-telling marketing campaigns on social media
6	Marketing, Strategic Partnerships	Business Travel	Engage partnerships with hotels frequented by business travels and consulting companies for batch discounts. Run aggressive story-telling campaigns on social media
7	Airport Management	Gate Location	Review gate scheduling algorithms used by airport, build capacity models and optimize gate scheduling algorithms to maximize the reduction of passenger delay
8	Digital User (Journey)	Online Support	Review online support KPIs, support technology and user

			experience. Optimize user experience and reduce MTTR.
9	Digital User (Journey)	App Boarding	Streamline mobile app user experience to increase signal-to-noise in alerting for boarding

Table 4: Business recommendations

Below we recall the business opportunity landscape that was initially considered and then overlay the areas covered by the recommendations (outlined in red) above and which additional areas may be pursued by future analytical studies.

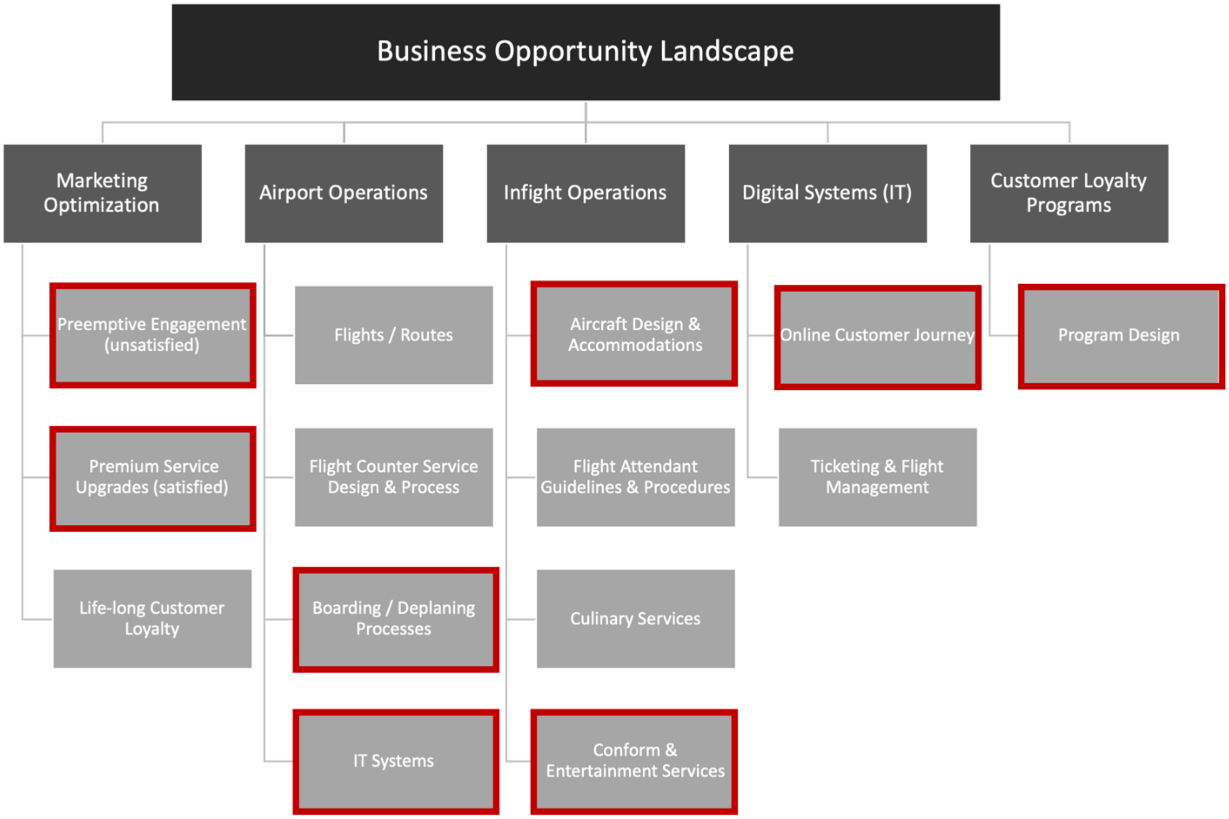

Business Opportunity Landscape
Marketing Optimization
Airport Operations
Infight Operations
Digital Systems (IT)
Customer Loyalty Programs
Preemptive Engagement (unsatisfied)
Premium Service Upgrades (satisfied)
Life-long Customer Loyalty
Flights / Routes
Flight Counter Service Design & Process
Boarding / Deplaning Processes
IT Systems
Aircraft Design & Accommodations
Flight Attendant Guidelines & Procedures
Culinary Services
Conform & Entertainment Services
Online Customer Journey
Ticketing & Flight Management
Program Design

Figure 25: Business opportunities addressed by this study

APPENDIX

FEATURE DICTIONARY

The below table shows all input features to the modeling process and of what type they are:

Feature Definitions for modeling (zscores)				
1	SATISFACTION	Sentiment attribute	float64	If is customer is satisfied with their travel experience
2	FEMALE	Customer attribute	float64	If the gender is noted as female
3	AGE	Customer attribute	float64	Age of the customer
4	LOYALTY_SCORE	Customer attribute	float64	Loyalty program member (yes/no)
5	BUSINESS_TRAVEL	Flight attribute	float64	If travel was designated for business or personal
6	SERVICE_CLASS	Flight attribute	float64	Service class of the ticket purchased e.g., economy, business, eco+
7	FLIGHT_DISTANCE	Flight attribute	float64	Miles between departure airport and the destination airport
8	DEPARTURE_DELAY	Flight attribute	float64	Minutes delayed in departing
9	APP_BOARDING_SCORE	Sentiment attribute	float64	Survey score for app boarding
10	ONLINE_BOOKING_EASE_SCORE	Sentiment attribute	float64	Survey score for online booking ease
11	CHECKIN_SERVICE_SCORE	Sentiment attribute	float64	Survey score for checking service
12	LEG_ROOM_SCORE	Sentiment attribute	float64	Survey score for leg room on plane
13	INFLIGHT_CLEANLINESS_SCORE	Sentiment attribute	float64	Survey score for inflight cleanliness
14	BAGGAGE_HANDLING_SCORE	Sentiment attribute	float64	Survey score for baggage handling
15	SEAT_COMFORT_SCORE	Sentiment attribute	float64	Survey score for seat comfort
16	FLIGHT_TIME_CONVENIENCE_SCORE	Sentiment attribute	float64	Survey score for flight time convenience
17	INFLIGHT_WIFI_SERVICE_SCORE	Sentiment attribute	float64	Survey score for inflight wifi service
18	FOOD_AND_DRINK_SCORE	Sentiment attribute	float64	Survey score for food and drink
19	INFLIGHT_ENTERTAINMENT_SCORE	Sentiment attribute	float64	Survey score for inflight entertainment

20	INFLIGHT_SERVICE_SCORE	Sentiment attribute	float64	Survey score for inflight service
21	ONLINE_SUPPORT_SCORE	Sentiment attribute	float64	Survey score for online support
22	GATE_LOCATION_SCORE	Sentiment attribute	float64	Survey score for gate location

Table 5: Feature dictionary

Websites

[1]. By Burak Himmetoglu. Stacking Models for Improved Predictions. Feb, 2017, https://www.kdnuggets.com/2017/02/stacking-models-imropved-predictions.html

[2]. By @pawangfg | Hyperparameters Optimization methods – ML. Jun 21, 2020, https://www.geeksforgeeks.org/hyperparameters-optimization-methods-ml/