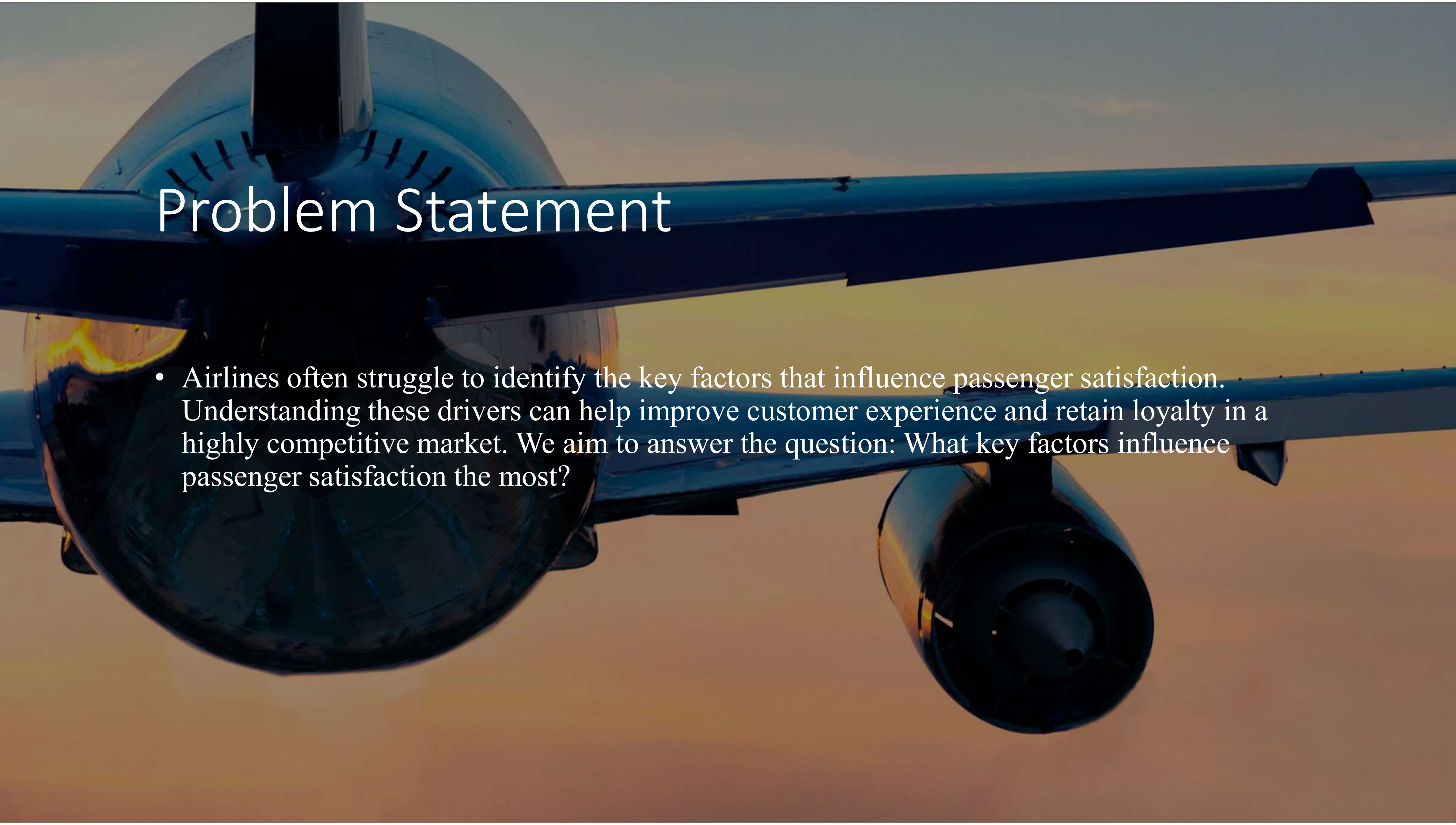
What Drives Airline Passenger Satisfaction?

Micro-Project #1

https://github.com/KcRyan7487/ANA500

Kasey Ryan

2025_May



Problem Statement

- Airlines often struggle to identify the key factors that influence passenger satisfaction. Understanding these drivers can help improve customer experience and retain loyalty in a highly competitive market. We aim to answer the question: What key factors influence passenger satisfaction the most?

Hypothesis Formulation

- Our hypothesis is that certain variables in the dataset will serve as strong and statistically significant predictors of passenger satisfaction, while others may not show a meaningful relationship. Specifically, we anticipate that the following variables will be among the strongest predictors:
  - Class
  - On-board service
  - Inflight service
  - Arrival Delay in Minutes

Acquire

- Dataset: airline.csv
- Description: The purpose of this file appears to be predicting target variable of "satisfaction" based on the other variables present in the dataset. A longer more description name might be something like "Airline Passenger Satisfaction". A slightly flashier name might be something like "What Drives Airline Passenger Satisfaction" which we have chosen as our official name for this project/analysis/paper/presentation.
- High Level Dataset metrics:
- Source: Retrieved from the provided list of datasets/.csv's to use as part of course materials for ANA-500
- Raw data: 129,880 rows, 25 columns

Prepare

Column	Populated Values	Missing Values	Distinct Values	Data Type
Gender	129880	0	2	object
Customer Type	129880	0	2	object
Age	129880	0	75	int64
Type of Travel	129880	0	2	object
Class	129880	0	3	object
Flight Distance	129880	0	3821	int64
Inflight wifi service	129880	0	6	int64
Departure/Arrival time convenient	129880	0	6	int64
Ease of Online booking	129880	0	6	int64
Gate location	129880	0	6	int64
Food and drink	129880	0	6	int64
Online boarding	129880	0	6	int64
Seat comfort	129880	0	6	int64
Inflight entertainment	129880	0	6	int64
On-board service	129880	0	6	int64
Leg room service	129880	0	6	int64
Baggage handling	129880	0	5	int64
Checkin service	129880	0	6	int64
Inflight service	129880	0	6	int64
Cleanliness	129880	0	6	int64
Departure Delay in Minutes	129880	0	466	int64
Arrival Delay in Minutes	129487	393	472	float64
satisfaction	129880	0	2	object

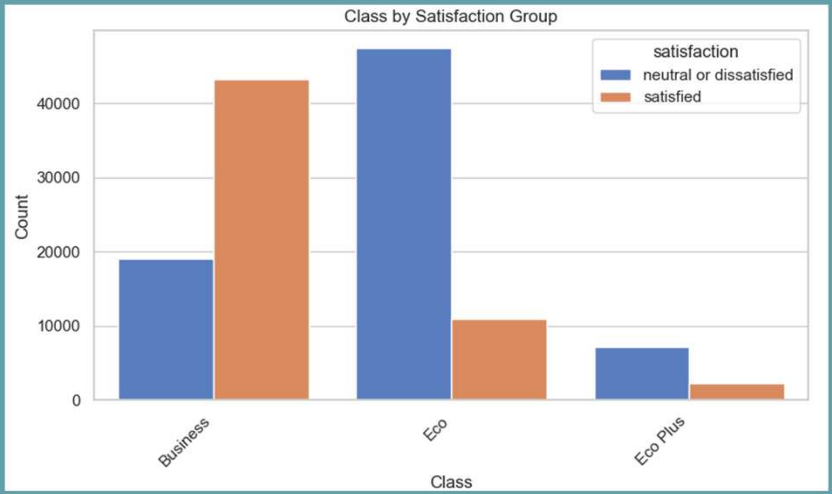
- Initial transformations: Drop the blank column at the beginning of the file (hard coded with name "Unnamed: 0“. Also drop the “id” column.
- Initial data integrity checks:
  - 0 duplicate rows
  - 393 missing values for “Arrival Delay in Minutes”

Prepare (additional considerations)

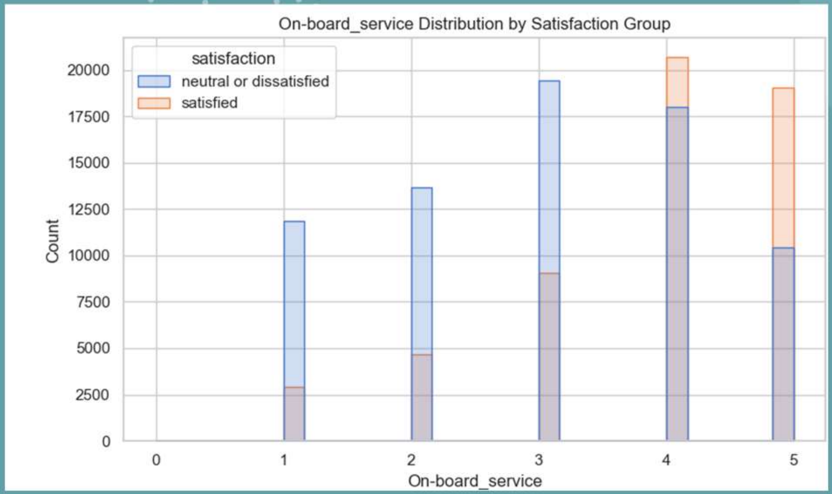
- Some of the variables have quite a few distinct values. It may behoove us to categorize these to a higher aggregation/granularity: Age, Flight Distance, Departure Delay in Minutes, and Arrival Delay in Minutes
- We'll have to decide what to do with the blank values for Arrival Delay in Minutes. We suspect this may be a strong predictor of satisfaction and so, if initial modeling supports that hypothesis and we suspect the variable should remain in the final model, it'll be very important we align on the best approach. Perhaps multiple approaches could be compared such as:
  - Dropping missing values
  - Imputing them based on the global average,
  - Imputing them based on various nearest neighbor methods
- Our target variable satisfaction only contains two possible values, a combination of neutral or dissatisfied or satisfied. Since our desired outcome is satisfied, we could recode these to satisfied as a 1 else 0 for ease of processing in future steps if/when needed.

Analyze data (EDA)

Exploratory Data Analysis was performed, examining distributions/frequencies for all variables and subset comparisons by the two satisfaction groups. Some hypothesized variables' examples are shown here; more information available in the supporting knitted pdf of the full analysis



- Business class passengers are overwhelmingly more likely to report satisfaction
- Economy class passengers are more likely to be dissatisfied or neutral
- Eco Plus has fewer total passengers overall, but also leans more toward dissatisfaction
- Suggests travel class may be strongly associated with satisfaction and is likely a key predictor



- Passengers who gave low ratings (1–2) for on-board service are far less likely to report satisfaction, while ratings of 4 and 5 show a sharp increase
- This visual implies a strong positive association between perceived on-board service quality and overall satisfaction.
- Suggests on-board service may be strongly associated with satisfaction and is likely also a key predictor.
- Further investigation needed for what specific components of the on-board service are considered in this rating to make this more actionable

Analyze data (Missing Values)

- We opted to drop the observations with missing values for our model building for now, though will be gathering feedback from stakeholders regarding similar analyses conducted using other imputation methods
  - We suggest proceeding with dropping because of the low volumes involved (393 out of 129,880 i.e. approx. 0.3%) and with it the advantage of our model being based upon truly observed data
- 129,487 observations remain, ready for use with logistic regression model(s)
  - 80/20 training/testing split
  - 103,589 training / 25,898 testing

Models Applied in this Analysis

- **Logistic Regression**
  - A classic statistical model for binary classification
  - Interpretable coefficients and fast to train
  - Works well when the relationship between features and the log-odds of the outcome is approximately linear
- **Support Vector Machine (SVM)**
  - A powerful machine learning algorithm for classification tasks
  - Finds the optimal boundary that maximizes the margin between classes
  - Performs well in high-dimensional spaces and can capture complex relationships with kernel tricks
- **Deep Learning (Recurrent Neural Network)**
  - Designed for learning sequential or time-dependent patterns
  - Can model complex nonlinear relationships and interactions across many features
  - Will be tested for potential improvements in predictive accuracy compared to the models above

Model Comparison Metrics

Before evaluating model performance, a quick debrief on "goodness of fit" metrics:

- Accuracy: Overall proportion of correctly predicted cases.
- Precision: Of all cases predicted to be positive (satisfied), how many actually were?
- Recall: Of all actual positive cases (satisfied), how many did the model catch?
- F1 Score: A balanced measure that combines Precision and Recall. Our pick for the most appropriate measure for this exercise of evaluation across different model types
- AIC (Akaike Information Criterion): Used to balance model fit versus complexity only for linear/logistic regression models (lower = better)
- MSE, MAE, and R-squared: Good for regressions; traditionally applied more towards continuous predictions vs our classification use-case thus not heavily utilized here

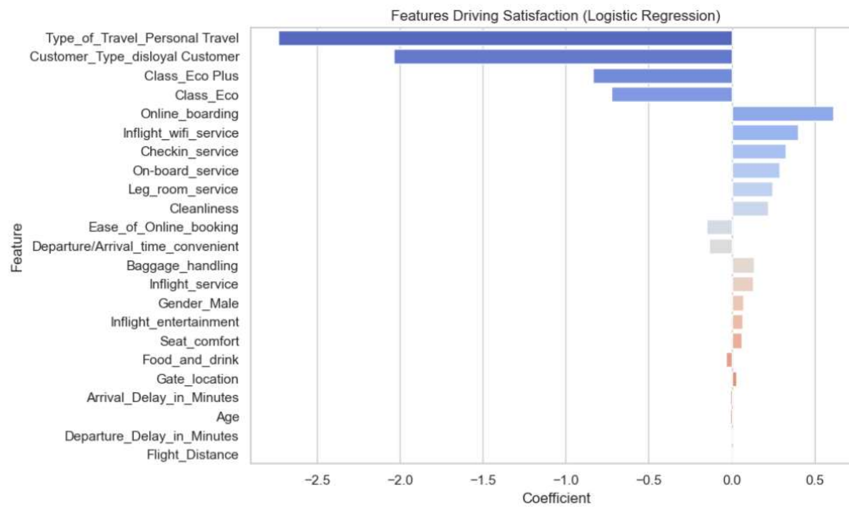
Analyze data (Comparing Logistic Regression Models)

- Full/complex model and other logistic regression models run to balance “goodness of fit” with simplicity and interpretability
- Comparing the results using the AIC* of:
  - Full/Complex Model (AIC= 69425.21)
    - All eligible variables included
  - Hypothesized Model (AIC = 107019.54
    - Including only the 4 variables in our original hypothesis (Class, On-board service, Inflight service, Arrival Delay in Minutes)
  - Most Parsimonious Model using <=4 variables (AIC = 93996.83)
    - Included Flight Distance, Online Boarding, Inflight Entertainment, Leg-Room Service
    - Lowest AIC among all combinations of <=4 variables suggests this is the better “simple” model, however AIC is still not better than the full model

*Akaike Information Criterion (AIC) rewards goodness of fit while “penalizing” complexity $AIC=2k-2\ln(L)$

Report

- Some features found to be most influential are shown here, more information available in the supporting knitted pdf of the full analysis
- Some hypothesized variables do appear to have a statistically significant relationship with satisfaction
- Others such as type of travel, customer type, and the “Online_boarding” variable appear to have a much stronger relationship as well
- Solid performance on the testing subset of data, BUT complexity of the full/complex model may still be an issue
    - 87% accuracy, 86% precision, 84% recall
- In future iterations further refinements are planned to balance explanatory power with reduced complexity/increased interpretability



Report (Logistic Model Comparisons)

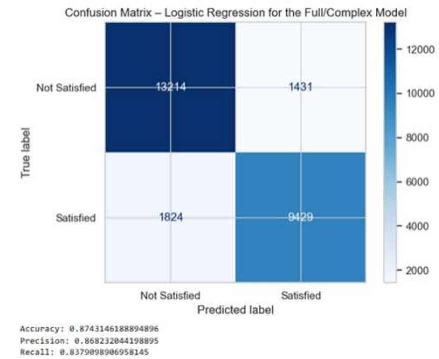
Full/complex model has solid performance on the testing subset of data, suggesting the model *may* not technically be overfit. As seen in the confusion matrix visual here:

- The full model correctly predicted 87% of cases overall
- High precision (86%) suggests most predicted satisfactions were truly satisfied and high recall (84%) suggests the model captured most actual satisfactions
- Forward stepwise selection to automatically optimize variable selection for best AIC also confirmed this model
- However, practical concerns of model complexity and interpretability remain, and thus, we experimented with pairing down the model's selected features to a smaller subset which still explains most of the variance

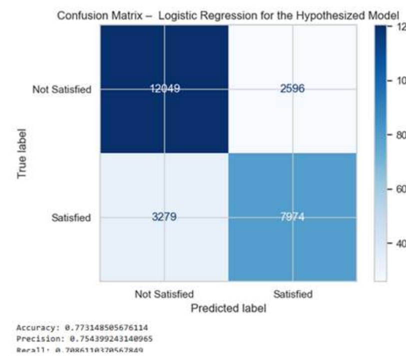
In further iterations we examined the AIC across 2 other logistic regression models, in an attempt to identify the best balance between complexity (i.e. explanatory variables used) and interpretability

- Hypothesized model
  - Class, On-board service, Inflight service, Arrival Delay in Minutes
  - Worse metrics across the board, though still had adequate explanatory power
- "Parsimonious Model" best AIC with <=4 variables)
  - "Brute forced" AIC analysis of all combinations of models with <=4 variables
  - Flight Distance, Online Boarding, Inflight Entertainment, Leg-Room Service
  - Better metrics versus the original hypothesized model but worse than full model

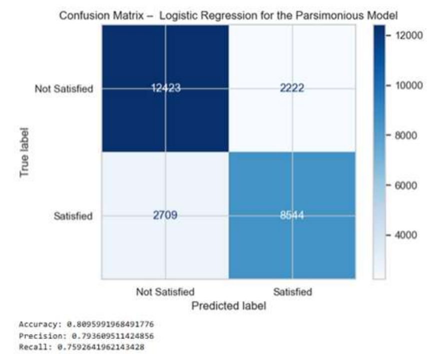
Full/Complex Model



Hypothesized Model



Parsimonious Model



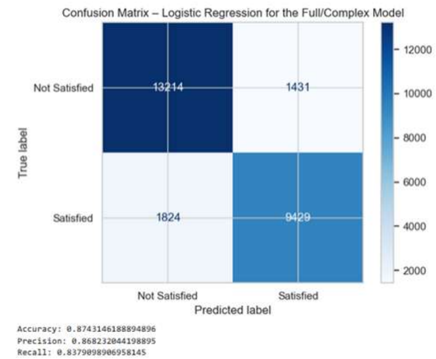
Model	AIC	Accuracy	Precision	Recall
Hypothesized	107019.5	77.3	75.4	70.8
Full/Complex	86715.67	87.4	86.8	83.8
Parsimonious (<=4 variables)	93996.83	81	79.4	75.9

Report (Logistic Model Comparisons) Cont

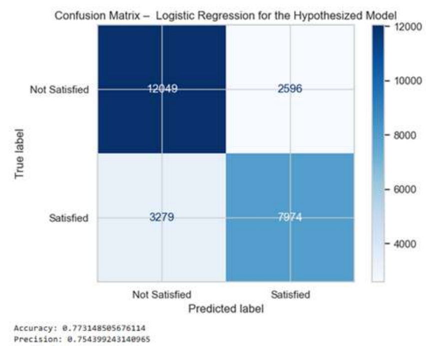
- Inclusion of Flight distance is particularly interesting and worthy of further investigation and consideration in future analyses.
  - Likely this suggests flight distance may act as a proxy for several quality-related features which are known to improve on long distance (especially international) flights, such as better onboard services, legroom, etc..
  - The importance increases when those correlated features are excluded which highlights the impact of multicollinearity in model interpretations
  - **i.e. flight distance is likely a proxy for better in-flight experience, not increased satisfaction from sitting longer**

- Ultimately the full/complex model would still be considered better, even utilizing AIC to penalize the complexity.
  - The <=4 variable parsimonious model may be a sufficient balance for our use-case, but this decision requires more input and consideration from our stakeholders and business domain experts
  - **The remainder of the analysis will utilize the full/complex model for comparisons of logistic regression versus the forthcoming SVC and Deep Learning model types**

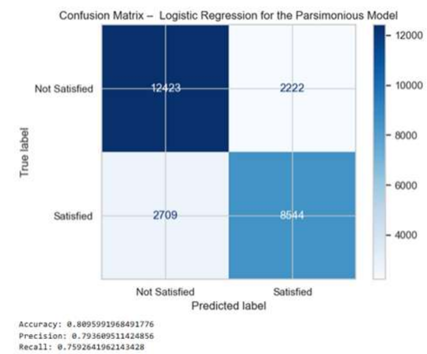
Full/Complex Model



Hypothesized Model

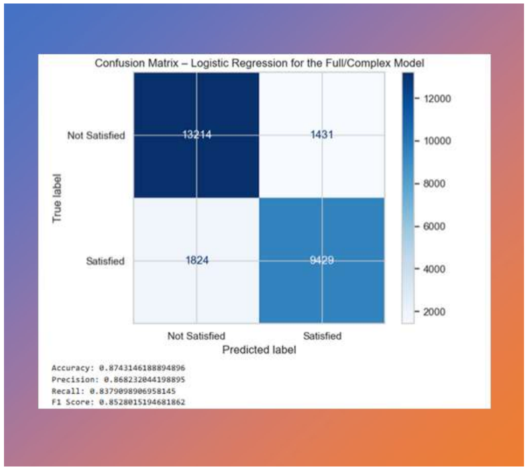


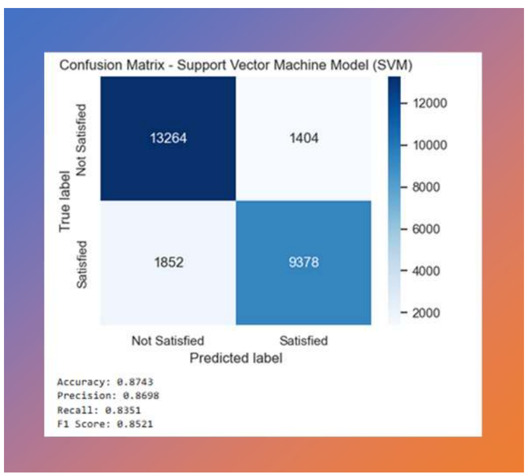
Parsimonious Model



Model	AIC	Accuracy	Precision	Recall
Hypothesized	107019.5	77.3	75.4	70.8
Full/Complex	86715.67	87.4	86.8	83.8
Parsimonious (<=4 variables)	93996.83	81	79.4	75.9

Report (SVM Model)



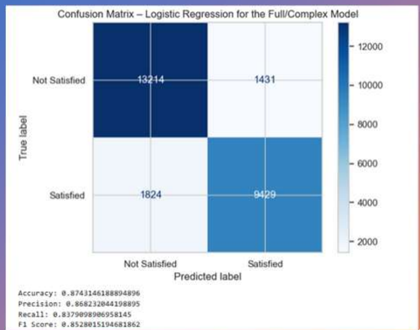


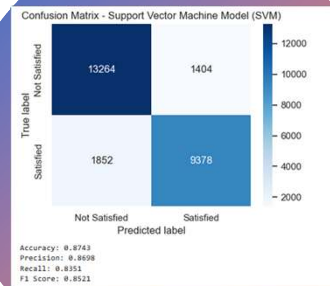
Model	Accuracy	Precision	Recall	F1
Full/Complex Logistic Regression	87.43146	86.8232	83.79099	85.28015
Support Vector Machine	87.43	86.98	83.51	85.21

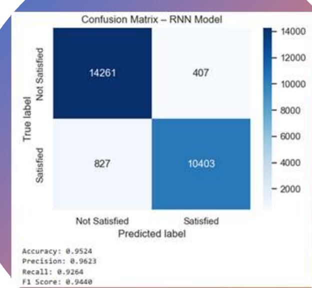
- Support Vector Machine (SVM) Model achieved accuracy of 87.4%, precision of 87.0%, recall of 83.5%, and F1 Score of 85.2%
- This is compared to the full/complex logistic regression model which achieved accuracy of 87.4%, precision of 86.8%, recall of 83.8%, and F1 Score of 85.3%
- Both models perform very similarly, but logistic regression slightly edges out SVM on F1 score and overall fit.
  - SVC had slightly higher precision by a small margin, but the rest of the metrics stood slightly worse
- Though LR could theoretically still be preferred for its simplicity and interpretability, especially in stakeholder-facing contexts.
  - Somewhat counteracted by the best performing LR model necessitating the full feature-set however; i.e. the full/complex LR model is still not exceptionally simple/interpretable

Report (RNN Model)

- Deep Learning Recurrent Neural Network (RNN) Model achieved accuracy of 95.1%, precision of 96.1%, recall of 92.4%, and F1 Score of 94.2%.
- This is compared to the full/complex logistic regression model which achieved accuracy of 87.4%, precision of 86.8%, recall of 83.8%, and F1 Score of 85.3%, as well as the SVM model which achieved accuracy of 87.4%, precision of 87.0%, recall of 83.5%, and F1 Score of 85.2%
- **The RNN model demonstrates the best performance across all models tested, with strong generalization to unseen data based on the held-out test set**







Model	Accuracy	Precision	Recall	F1
Full/Complex Logistic Regression	87.43146	86.8232	83.79099	85.28015
Support Vector Machine	87.43	86.98	83.51	85.21
Deep Learning Recurrent Neural Network	95.05	96.07	92.37	94.18

Act

- Our original hypothesis proposed that Class, On-board Service, Inflight Service, and Arrival Delay (min) would be the strongest predictors of passenger satisfaction.
- Results suggest the story is more nuanced, as several other variables showed strong predictive power, and the ranking of importance varied by model.
- While both logistic regression and SVM offer interpretable models, our highest-performing model was the Recurrent Neural Network. These are inherently less transparent, making it difficult to trace individual feature contributions.
- This "black box" nature of deep learning raises important ethical and practical considerations, particularly for high-stakes business decisions.
- Despite that, the RNN model’s superior performance indicates that nonlinear relationships and/or interactions between many variables likely drive satisfaction in ways not captured by simpler models.
- For actionable insight, we recommend our stakeholders and business domain experts balance accuracy with interpretability, and consider using interpretable models for communication and diagnostics, while leveraging deep learning for behind-the-scenes prediction where appropriate.
- These results reinforce that while data can guide decisions, thoughtful interpretation and domain expertise remain essential to turning model results and predictions into meaningful action.