Project Overview: Customer Churn in the Airline Industry

Southeast Airlines needed to lower their customer churn (sometimes referred to as customer attrition). Like many airlines, Southeast has, up until now, believed that the best way to minimize customer churn was to have a robust loyalty program for frequent flyers. The basics of Southeast's loyalty program were similar to other airlines. In short, the airline rewarded repeat business and as a customer flew more, they would rise through levels of service and "bank" miles that could be redeemed for free or discounted travel. There was no data to back up this line of thinking, however, as it was just the "accepted industry best practice". However, their customers were valuing the loyalty program less, which was one reason why just relying on their loyalty program might not be sufficient in keeping low customer churn. In fact, according to a recent International Air Transport Association (IATA) study, airlines carry $12B in "loyalty debt" and frequent flier mileage and points are slowly devaluing, while the overall balance (or debt) is increasing.

Additionally, customer churn is actually a lagging indicator, meaning the loss has already occurred. As such, it was a measurement of the damage inflicted. The real goal is to reduce churn by getting ahead of the loss (of the customer) by identifying some leading indicators, or metrics, that might help keep a customer. In other words, these leading indicators, or metrics, could help identify when a customer was about to stop flying Southeast. These insights could provide actionable suggestions as to how to avoid having the customers leave and go to another airline. In thinking about customer churn, several key facts are relevant:

Net Promotor Score (NPS): NPS asks customers to respond, on a scale of 1–10, to one simple question: "How likely is it that you will recommend our airline to a friend or colleague?". If respondents score less than 7, they're detractors. If they scored above an 8, they're promoters. In the middle range (a score of 7 or 8), then they're "passive". In a given group, subtracting the percent of respondents who are detractors from the percent of respondents who were promoters provides the overall NPS score. The concept of NPS is that customers who are promoters are good customers to keep. Such customers may sometimes even provide free "word of mouth" advertising. Customers who are detractors are really problematic in that they may actively tell their social connections not to use the product or service (i.e., they would be telling people not to fly Southeast). It has often been suggested that NPS provides a good proxy for understanding how likely a customer is to churn. For example, according to one source, NPS is nearly three times more sensitive at predicting customer churn than customer satisfaction. In addition, detractors are 1.5 times more likely to stop using a service as compared to promoters. In short, analytics could be a key to the success of Southeast.

Southeast and its Regional Airline Partners: The airline had many regional airline partners that operated quasi-independently. Southeast Airlines is one of the top four airlines in the United States. Like the other large airlines in the U.S., customers buying a Southeast plane ticket fly on Southeast Airlines primary routes as well as on Southeast's regional partner airlines. Regional airlines act as feeder airlines to major airlines by connecting smaller airports to the airline's main hubs. Hub airports are always located in major cities, whereas the regionals serve smaller cities and rural areas. Like other airlines, Southeast contracted out to regional carriers because it allowed them to lower their risks related to capacity and pricing. Specifically, regional airline contracts last for a number of years, after which, Southeast can renegotiate to adjust (up or down) the number of flights provided by that partner. This enables Southeast to more easily reflect their current market conditions. It is possible, for example, that if demand falls, Southeast would not renew some of their regional contracts. On the other hand, if demand rises, Southeast can expand their contract, and bring more planes into service more quickly than they could on their own. Note that NPS was not currently used as part of Southeast's partner airline strategy (i.e., it was not part of the data Southeast used to help decide which partners to keep, which partners to drop and which regional airlines should become new partners).

The Data Available: Southeast often surveyed their customers, and in fact, possessed thousands of recently completed customer surveys. Southeast has been using the surveys to calculate NPS. They would increase their focus on providing good customer service when their NPS score went down. This was typically via a memo to customer facing staff, where they were encouraged to "smile more". The survey dataset contained thousands of observations of flight segment data collected by Southeast Airlines. Each row represents one flight segment, by one airline (either southeast or one of its partner airlines), for a specific customer. Each column represents an attribute of that particular flight segment. Each row captures 26 characteristics of the flight (ex. day of month, date, airline, origin and destination city, if the flight was delayed), the customer (ex. age, gender, price sensitivity, the person's frequent flyer status). The row also contains a simple survey-based rating of each customer's likelihood to recommend the airline that they just flew as well as a field for open-ended text comments. It should be noted that there are some missing values in the dataset. The table below provides a short description for each attribute.

Attributes:

1. ***Likelihood to Recommend –*** *rated on a scale of 1 to 10, which shows how likely the customer is to recommend the airline to their friends (10 is very likely, and 1 is not very likely).*
2. ***Airline Flyer Status*** *– each customer has a different type of airline status, which are platinum, gold, silver, and blue (based on level of travel with the airline)*
3. ***Age*** *– the specific customer's age. Ranging from 15 to 85 years old.*
4. ***Gender*** *– male or female.*
5. ***Price Sensitivity*** *– the grade to which the price affects to customers purchasing. The price sensitivity has a range from 0 to 5.*
6. ***Year of First Flight*** *– this attribute shows the first flight of each single customer. The range of year of the first flight for each customer has been started in 2003 until 2012.*
7. ***Flights Per Year*** *– The number of flights that each customer has taken in the most recent 12 months. The range starting from 0 to 100.*
8. ***Loyalty*** *– An index of loyalty ranging from -1 to 1 that reflects the proportion of flights taken on other airlines versus flights taken on this airline. A higher index means more loyalty.*
9. ***Type of Travel*** *– One of business travel, mileage tickets, or personal travel (ex. vacation)*
10. ***Total Frequent Flyer Accounts*** *– How many frequent flyer accounts the customer has.*
11. ***Shopping Amount at Airport*** *– The spending on non-food & services at the airport (in $)*
12. ***Eating and Drinking at Airport*** *– The spending on food/drink at the airport (in $).*
13. ***Class*** *– three different kinds of service level (business, economy plus, and economy).*
14. ***Day of Month*** *–the traveling day of each costumer (ranges from 1 to 31).*
15. ***Flight date*** *– the passenger's flight date of travel.*
16. ***Partner Code*** *– This airline works with wholly- and partially-owned subsidiary companies to deliver regional flights. For example, AA, AS, B6, and DL.*
17. ***Partner Name*** *– These are the full names of the partner airline companies.*
18. ***Origin City*** *– the place where passenger departed from. For example, Boston MA.*
19. ***Origin State*** *– the place where passenger departed from. For example, Texas.*
20. ***Destination City*** *– the place to which passenger travels to. For example, Boston MA.*
21. ***Destination State*** *– the place to which passenger travels to. For example, Texas.*
22. ***Scheduled Departure Hour*** *– the specific time at which the plane was scheduled to depart.*
23. ***Departure Delay in Minutes*** *– How long the flight's departure was delayed, when compared to schedule.*
24. ***Arrival Delay in Minutes*** *– How long the arrival was delayed.*
25. ***Flight Cancelled*** *– occurs when the airline does not operate the flight.*
26. ***Flight time in minutes*** *–the length of time, in minutes, to reach the destination.*
27. ***Flight Distance*** *– the distance between the departure and arrival destination.*
28. ***Origin Longitude*** *– longitude of origin city*
29. ***Origin Latitude*** *– latitude of origin city*
30. ***Destination Longitude*** *– longitude of destination city*
31. ***Destination Latitude*** *–latitude of destination city*

Example attributes that could be added include:

1. **Arrival Delay greater 5 Minutes** – It means the delay of arrival airline time, which is more than 5 minutes per each passenger in the data.
2. **Long Duration Trip** – A Boolean variable that divides flight segments into two types: FALSE means a shorter duration segment (including average delays), TRUE means a longer duration segment.

The overall goal of the case is to provide actionable insight, based on the data available.

There are three deliverables for this project:

Deliverable #1

One slide presentation (10-15 slides) that summarizes your analysis. The audience for this will be executive-level leadership of an airline company. Please assume these executives do not know too much about statistics, so you probably should not show R code or quote terms like "R-squared" or "p-value," but rather describe your statistical results in plain language. Your group will present this slide deck on the last day of classes at the normal class time (**Thursday, Dec. 3, 5 p.m. EST**).

Deliverable #2

One MS-Word file, containing a detailed report of all your work. This report should include sections for all the phases of data science discussed in this course. Your intended audience for this report will be your Data Science professors, who understand R code and Data Science. Please make sure to include all assumptions made and any analysis completed, whether you found it significant or not.

Deliverable #3

All your R Code: One .R file, well structured, clean, with lots of comments.

Rules of Engagement: This is an honor system assignment: You **may** consult with your other group members, IST687 professors and faculty assistants, the textbook, and publications on the Internet at any time. Your attribution statement, at the top of your R-code file, must reflect these constraints. As a group, you may not share your results or work in progress with any other human besides your professors and faculty assistants.

Project Goal: The goal of this term project is for you to use all of the skills you have developed in the IST687 class/labs/homework to make sense of a novel dataset; to perform some essential analyses on the dataset; and to explain/document what you have done. The dataset contains survey data of air travel within the U.S, one row per customer, per trip.

Accessing Your Data File: The data will be available to you in Blackboard. The file contains about 31 columns/variables.

Recommended Project Phases

Data Pre-processing / Data Preparation Phase

- *Phase 1: Clean your Data*. There are several columns in the dataset that may contain missing data, need cleaning, or transformation to a different data type, etc. Write code that examines each column to see if it contains missing (or NA) data. To mitigate missing data, use any of the methods learned in class and make sure to note your approach in your final documentation. Use comments in your code to document how many missing data values you had to repair and your approach to do that. Write code to clean your data, as well as any needed code to transform your data.

- *Phase 2: Getting to know your Data and prepare business questions*. For all of your significant numeric variables, create visualizations, tabular summaries, and comments that describe what value you gain from these. Describe in detail how this drove your business question development. Include a list of those business questions (~10), as well as the data segments you are using in each question.

Exploratory Analysis Phase

- *Phase 3: Predictive Modeling*. Many columns contain data relating to the characteristics of each customer's trip. To get further insights into your business questions, use the modeling techniques we learned in the class (Linear Modeling, Assoc Rules, SVM) or others you are familiar with in R, to develop 3-5 different predictive models that analyze the data. This is 3-5 models per modeling technique. This means you need to change parameters and/or model inputs to get the best model.

Business Recommendations Development Phase

- ***Phase 4****: Make sense of your models and Develop a Marketing Plan*. Use the results of your analysis and models to identify 3 customer segments in need of action (e.g., middle age, female, business travelers, in the northeast). Finally, for each segment, make 3 recommendations that are **actionable** and you believe would increase the NPS for the segment.

Hints:

1. For exploratory work:
    a. Histograms and boxplots of numeric variables are typically useful (except attributes such as olong, olat, dlong, and dlat).
    b. Producing tables of categorical response variables (e.g., Gender) if often helpful.
    c. Boxplots of Likelihood.to.recommend, using each of these grouping variables is often useful (e.g., grouping Gender or Type.of.Travel and then generating boxlots or histograms within each grouping).
    d. Barplots of NPS across different categories is often useful (e.g., showing NPS by Type.of.Travel)
2. Since there is lat and long data, you should make maps U.S. and overlay:
    a. Dots on the map where the color of the dot indicates the value of Likelihood.to.recommend.
    b. State level means for Likelihood.to.recommend
    c. Explore origins of the flight (rather than the destinations).
3. It might be helpful to create a new column that is Detractor (it could be Boolean) and try to understand rules for predicting a Detractor.
4. For your recommendations:
    a. Provide a summary covering all of your results in language that is suitable for a manager to understand. Most managers do not know too much about statistics, so you probably should not quote terms like “R-squared” or “p-value” but rather describe your results in plain language.
    b. **Important**: Your recommendations MUST be connected with one or more of your data science results; they MUST NOT be based on your own personal experience with flying, airlines, airports, etc.