K353 Project: Business Analytics at American Airlines

Background: Several new airports have opened, opening the market for new routes (a route refers to a pair of airports). American Airlines is evaluating its fare (ticket price) decisions on these new routes.

Southwest, a discount airline, is a competitor of American Airlines. Southwest's strategy (point-to-point routes covering only major cities, use of secondary airports, standardized fleet, low fares) has been very different from the model followed by the older and bigger airlines like American Airlines (hub-and-spoke model extending to even smaller cities, presence in primary airports, variety in fleet, pursuit of high-end business travelers). The presence of discount airlines is therefore believed to reduce the fares greatly. To stay competitive, American Airline needs to consider whether Southwest is serving on a route.

Dataset: American Airlines has provided you with a dataset called KnownRoutes.csv which contains real data that was collected for the third quarter of 1996. The table below summarizes the variables in the KnownRoutes.csv dataset. Note that some cities are served by more than one airport, and in those cases the airports are distinguished by their three-letter code.

Table 1 Documentation for the KnownRoutes.csv dataset

Code	Description
S_CODE	starting airport's code; do NOT use in analysis
S_CITY	starting city; do NOT use in analysis
E_CODE	ending airport's code; do NOT use in analysis
E_CITY	ending city; do NOT use in analysis
COUPON	average number of coupons (a one-coupon flight is a non-stop flight, a two-coupon flight is a one stop flight, etc.) for that route
NEW	number of new carriers entering that route between Q3 and Q2
VACATION	whether a vacation route (Yes) or not (No); Florida and Las Vegas routes are generally considered vacation routes
SW	whether Southwest Airlines serves that route or not (1 = Yes, 0 = No)
HI	Herfindel Index – measure of market concentration
S_INCOME	starting city's average personal income
E_INCOME	ending city's average personal income
S_POP	starting city's population
E_POP	ending city's population
SLOT	whether either endpoint airport is slot controlled or not (1 = Controlled, 0 = Not controlled); this is a measure of airport congestion
GATE	whether either endpoint airport has gate constraints or not; this is another measure of airport congestion
DISTANCE	distance between two endpoint airports in miles
PAX	number of passengers on that route during period of data collection
FARE (target variable)	average fare on that route

Objectives

1) **Data exploration and preparation:**
   - Execute business understanding, data understanding and data preparation steps. Summarize your findings and justify any modifications you made to the dataset.

- Explore the numerical and categorical predictors that can be used to predict the average fare on a route. Which predictors seems to be the best, and why?

2) **Building a predictive analytics model:**
   - Generate at least three predictive analytics models to be used to predict the average fare of a route. To demonstrate the breadth of your expertise, please plan to apply various techniques to build and improve the models you are proposing.
   - Provide a predictive performance comparison among all the models you investigated. Identify the best model and justify your selection.

3) **Building a prescriptive model:** Build a simulation model to help American Airlines understand the revenue that might be earned from a new route that is being planned. Your model should address the uncertainties about whether Southwest will serve on a route, what the average fare on that route will be, and the number of passengers who will travel on that route. American Airlines provided the following information to be used to build your simulation model:
   - The probability that Southwest will serve on this new route is 0.7.
   - The number of passengers is approximately normally distributed with mean 4700 and standard deviation of 500.
   - The market share of American Airlines, in other words the fraction of passengers that travel with American Airlines on this route, is expected to be 17.6%. If Southwest is not serving on a route, American Airlines does not expect market share to be impacted by the ticket fare. However, if Southwest is serving on a route, customer behavior is more price sensitive. American Airlines estimates that 20 customers are lost for each dollar charged over the average fare price and 20 customers are gained for charged below the average fare price.
   - The goal of American Airlines is to maximize their revenue. American Airlines is considering three options for setting its ticket fares on the new routes: 1) 90% of the average fare, 2) same as the average fare, or 3) 110% of the average fare. American Airlines would like your input on which pricing strategy is preferable on each route.

Using your simulation model, how would you compare the three pricing strategies American Airlines is considering according to the results from your simulation model?

4) **Your business analytics vision:** Provide a recommendation about other business analytics questions you can answer with your skill set that would help American Airlines improve their operations. As you describe your vision, please be specific with respect to the pros and cons of your suggestions and data that would be required to implement them.

Submission (100 Points)

Deliverables
- Report
- R script used for analysis

Expectations
The goal of the report is to provide a concise summary of the process undertaken in this project. Your report should answer all the questions outlined in the Objectives section and should be able to stand for itself in terms of logic, assumptions, process, and outcome. You do not need to include the plots you generated in the report if the R code to generate those plots are available.

Your R script should include all analysis you conducted for this project and support the analysis presented in the slide deck. It is important to keep in mind that your R script should be well-documented, well-commented and free of errors.

Grading Rubric for Final Submission

Category	Did not meet expectations	Fair	Good	Excellent
Data Understanding and Preparation (25%)	Concepts from the class are mis-applied. Some objectives are not addressed.	Concepts from the class are understood but may or may not be applied correctly. Some objectives are not addressed.	All objectives are addressed, and concepts are applied correctly.	All objectives are addressed. Makes extensive use of concepts learned in class. Shows due diligence in data exploration and preparation.
Predictive Analytics (25%)	Concepts from the class are mis-applied. Some objectives are not addressed.	Concepts from class are understood but may or may not be applied correctly. Some objectives are not addressed.	All objectives are addressed, and concepts are applied correctly.	All objectives are addressed. Makes extensive use of concepts learned in class. Shows due diligence in model creation and comparison.
Prescriptive Analytics (25%)	Concepts from the class are mis-applied. Some objectives are not addressed.	Concepts from the class are understood but may or may not be applied correctly. Some objectives are not addressed.	All objectives are addressed, and concepts are applied correctly.	All objectives are addressed. Makes extensive use of concepts learned in class. Shows due diligence in model creation and use of the model for decision making.
Files and Presentation (25%)	Files are lacking in formatting requirements, are incomplete, generates major errors, or for some other reason would be unacceptable to present to a management audience.	R file is generating errors, there are uncommented/unclear sections, or slide deck and presentation are unorganized, unclear, and hard to understand.	R file and slide deck are complete and logical with only minor mistakes (e.g., unclear sentences, formatting issues, unorganized), or there were minor issues with the presentation (e.g., unorganized, minor mistakes)	R file and slide deck are complete, well-organized, comprehensive, and easy to understand. R file does not generate any errors. Presentation is organized, engaging and professional.