Ryan Louie ‘17
Kyle McConnaughay ‘15
Data Science - Spring 2014
Franklin W. Olin College of Engineering

ICFI Project Final Report

Introduction to Problem Space

Airlines operate thousands of flights and transport millions of people every year. In order to maximize marginal revenue for each flight it is imperative to fill as many seats as possible and maximize the average price of ticket sold on each flight.

Airlines’ current revenue optimization policy is the Expected Marginal Seat Revenue b (EMSRb) strategy. EMSRb assumes that the number of purchases made in a given booking class are normally distributed (the actual distribution $N(\mu, \sigma)$ is taken from past booking data) and that customers who cannot buy a cheap ticket will ‘buy up’ into the next most expensive booking class. On the basis of these assumptions, EMSRb will throttle the number of cheap tickets available to customers so that the expected revenue from an expensive booking class will always be equal to or greater than the next most expensive booking class.

This is an old method developed in the 1990’s that makes a number of assumptions about consumer behavior and does not take advantage of advances in computing power and machine learning techniques. We ran many different kinds of machine learning algorithms on booking data to develop a better model for predicting bookings.

Data Management

Background

The data set provided by ICFI details every flight operated by an unknown airline during a three month period from January to March 2013. ICFI's Data Management Revenue Service collected it in order to visualize and predict demand for flight routes operated by the airline.

Legal and Ethical Issues

ICFI requires that person who work with this data sign a Non Disclosure Agreement (NDA). The terms of the NDA prevent disclosure of any part of the data set to the public for any reason. Code written to analyze the data set, however, may be put in the public domain.

There are no major ethical issues implicit with the use of this data set. Each record is anonymized and contains no information that could be used to identify individual customers.

Storage, Processing, and Sharing Plan

The data is kept in a two gigabyte CSV file. The terms of the NDA do not allow the data to be hosted on any third-party storage systems, but multiple, locally stored copies of the data set are permitted. Processing of the data can also happen locally, as the size of the data set is small enough to fit on a hard drive.

Exploration of Data

Terminology

Term	Explanation	Abbreviation
Authorization	*Authorization* is the number of bookings the airline company is willing to sell in a particular booking class. It is one key "knob" which the airline uses to control booking behavior.	AUTH
Authorization Curve	*Authorization curves* reflect the history of Authorization over the time of the booking period.	AUTH Curve
Booked	Number of seats currently booked in a particular booking class	BKD
Booking Class	BOOKING CLASS LETTER. EACH BOOKING IS MADE UNDER A CERTAIN BOOKING CLASS. THIS BOOKING CLASS DETERMINES THE FARE OR PRICE. THERE IS A PARTICULAR HIERARCHY OF BOOKING CLASSES	BC

Cabin	Either business or economy. Each booking class is associated with one and only one cabin.	CAB
Cabin Load Factor	(TOTALBKD / CAP) at departure	CLF
Capacity	Capacity of cabin in terms of seats.	CAP
Delta Booked	Change in BKD during time t = t0 and t= t1	ΔBKD
Date	Date of a flight's departure. Corresponds to keyday zero.	DATE
Destination	The airport at which a flight arrives	DEST
Keyday	Number of days before departure that a row of data was inserted.	KEYDAY
Origin	The airport from which a flight leaves	ORG
Total Booked	The number of bookings realized at departure in a booking class' cabin (either business or economy	TOTALBKD

Analysis

+ CDF OF CLF ACROSS ALL OF OUR DATA (DRILL DOWN Y?)

+ AND RELATIVE FREQUENCY OF TICKETS SOLD IN EACH BOOKING CLASS
-scatter plots

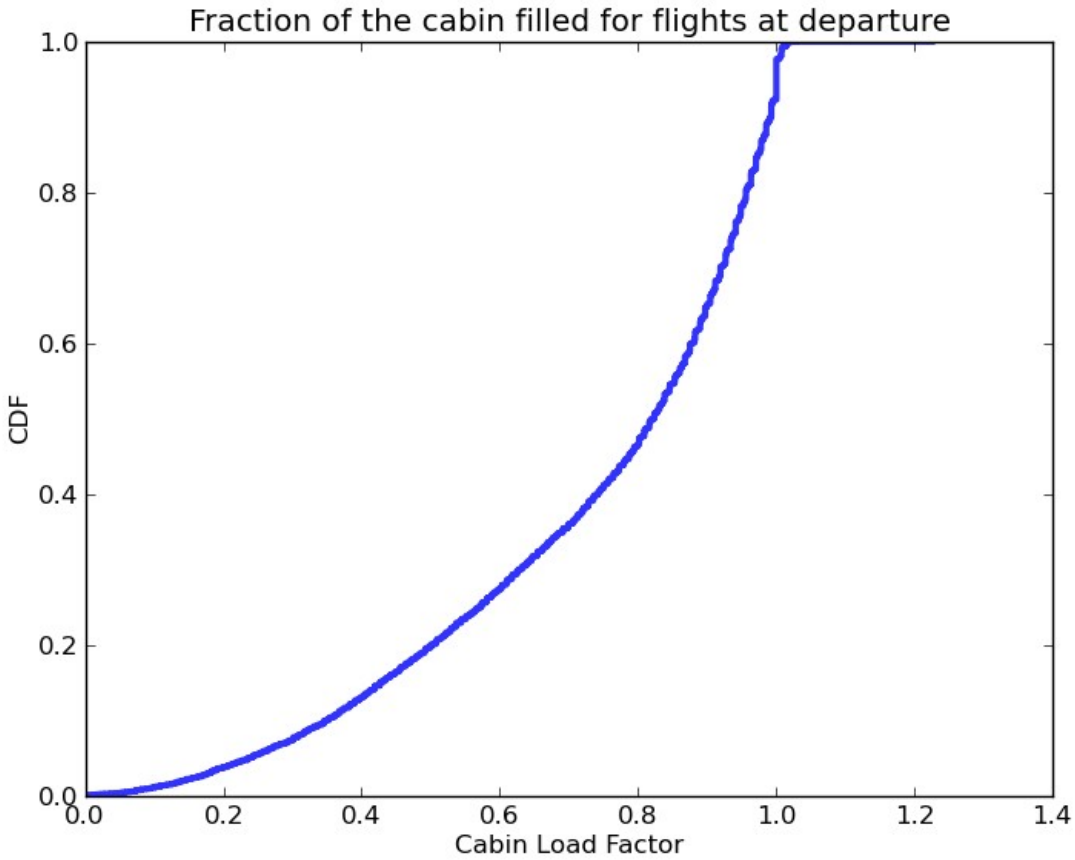


Figure 1: Roughly half of all flights are only at 85% capacity. The ideal cabin load factor is 95% - 100%.

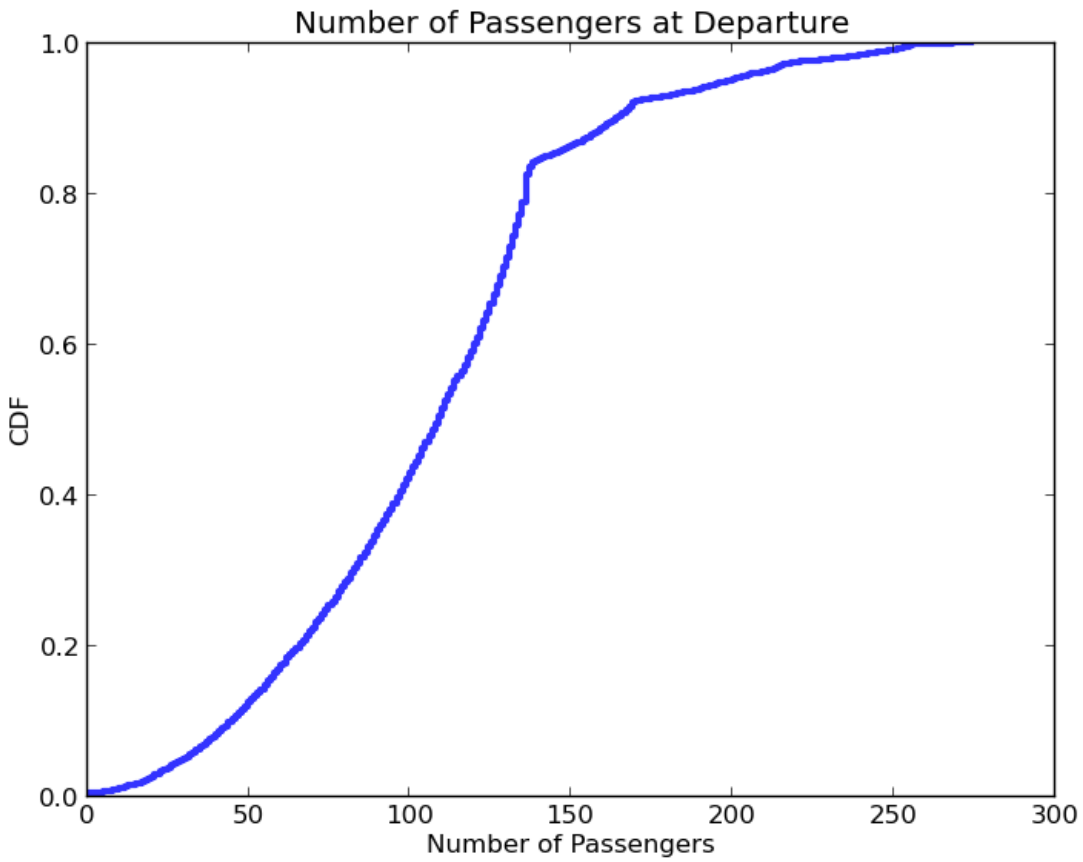


Figure 2: The airline owns planes of three different capacities; 140, 165, and 210. At each of these breakpoints there is a 'mini CDF' embedded in the larger one, and each mini CDF is similar in shape to the others. There are significant percentages of flights flying at less than optimal capacity.

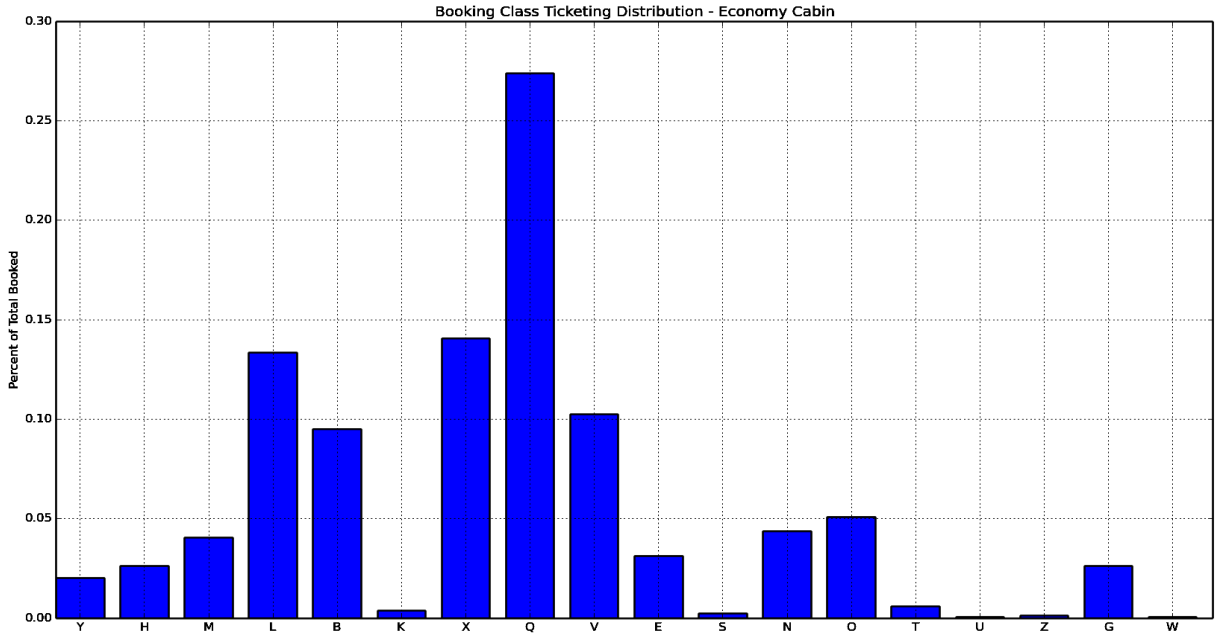


Figure 3: Most expensive to least expensive booking classes from left to right. Bar values represent the percentage contribution of a booking class to the total booked seats. There are a few heavily-utilized booking classes that are important features for learning.

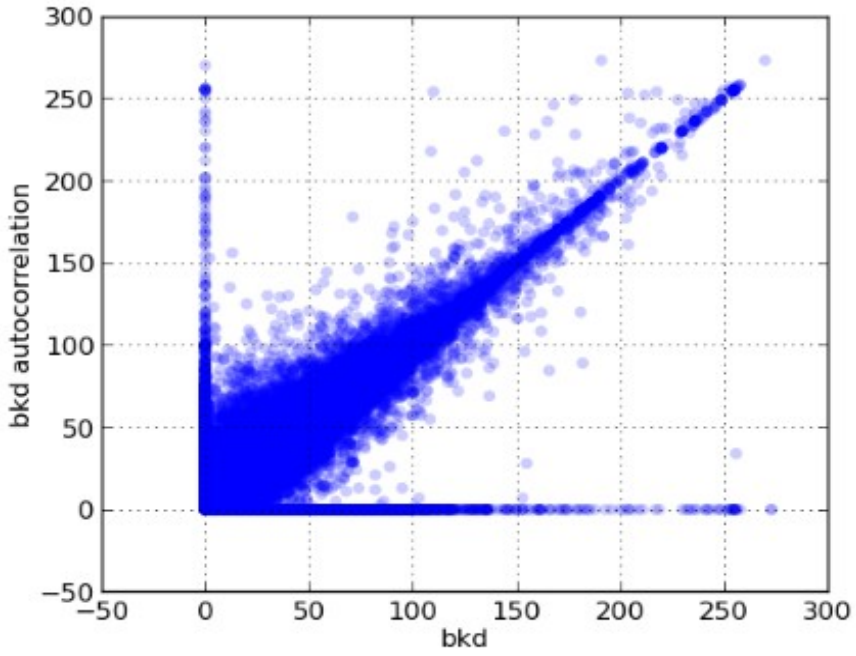


Figure 4: Positive cross-correlation between bkd at time t_i (X) and bkd at time t_{i+1} (Y). The line y = x tells the story of bookings that hardly change between timesteps, (I.e the booking behavior is stable.) This stability is more noticeably found in planes with 200 or more people booked.

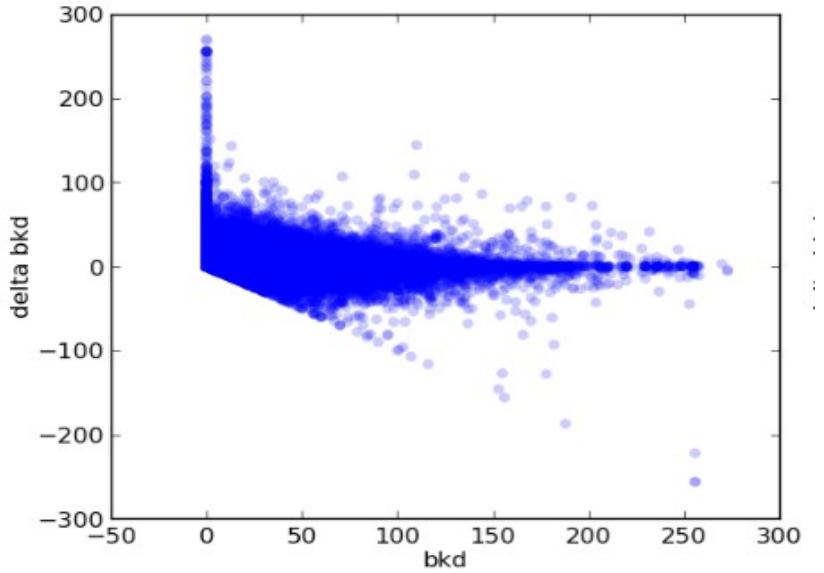


Figure 5: The magnitude of ΔBKD speaks to the volatility of consumer booking behavior. This is plotted against the amount of seats currently booked. As noted earlier, stability, or small volatility, is found when more seats are currently booked. The negatively sloping line y = -x describes the fact that a plane cannot have more cancellations than it has bookings.

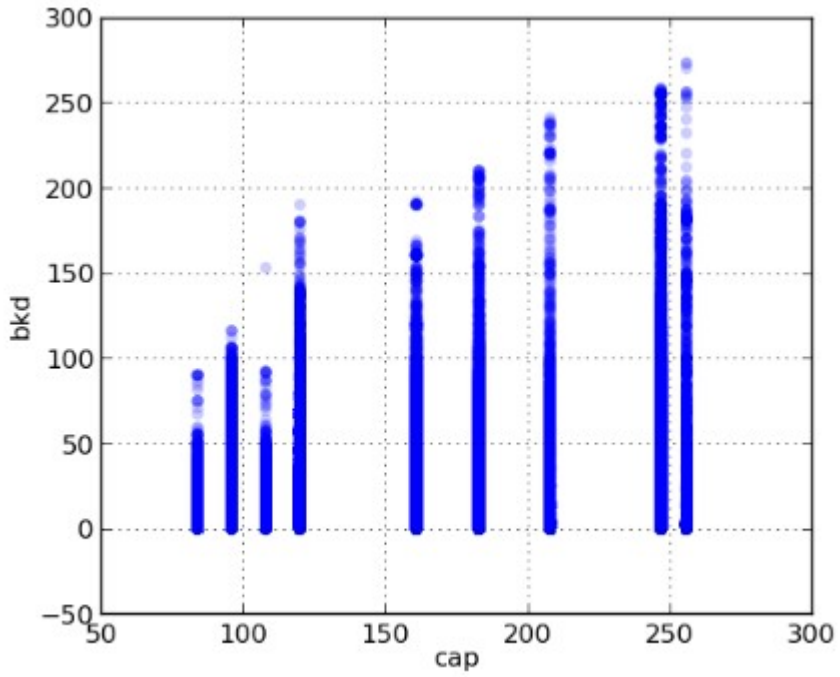


Figure 6: Capacity of the Cabin versus Total Booked for all flights. The positive correlation between the two indicates to our learner that, for a large flight, each ΔBKD will be larger on average than each ΔBKD for a smaller plane.

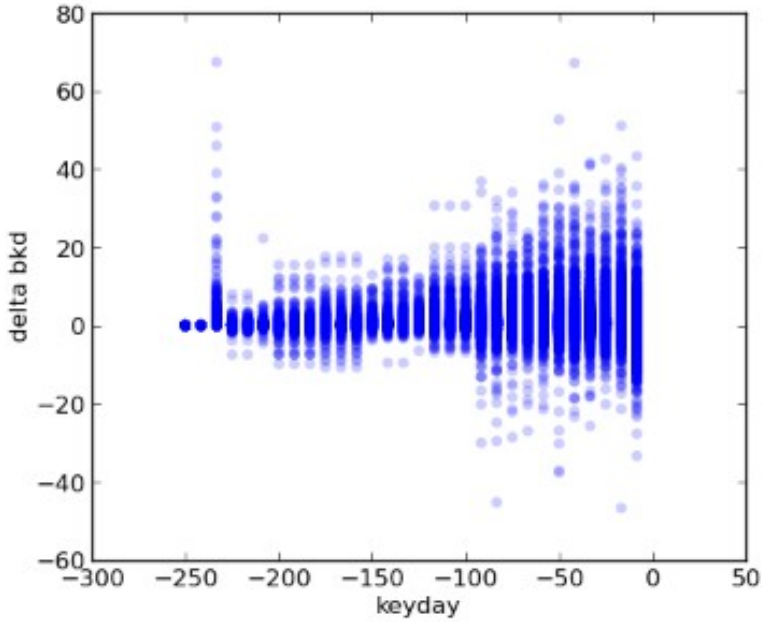
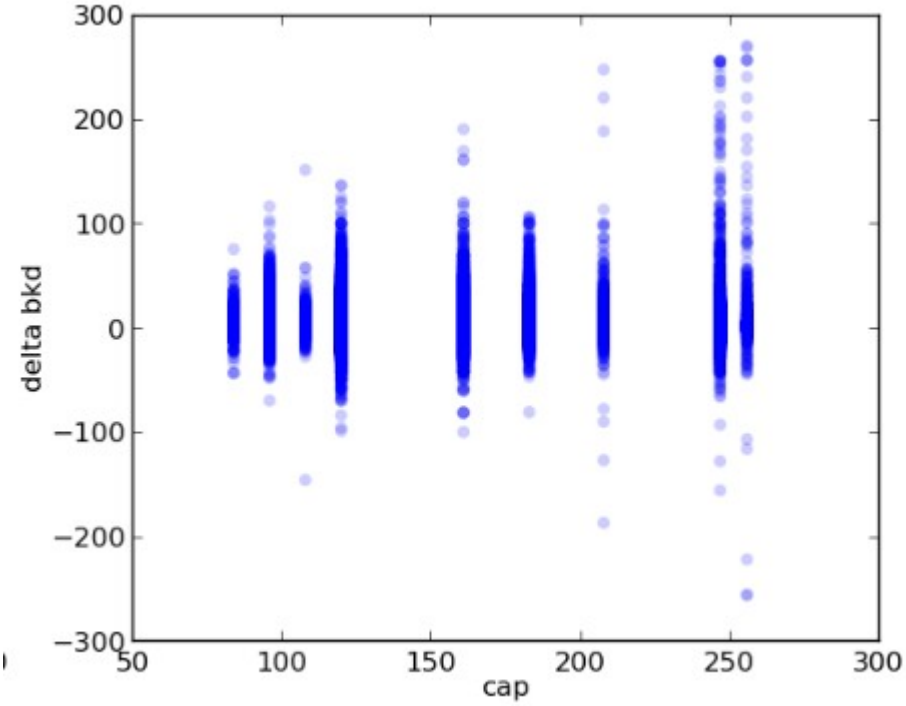
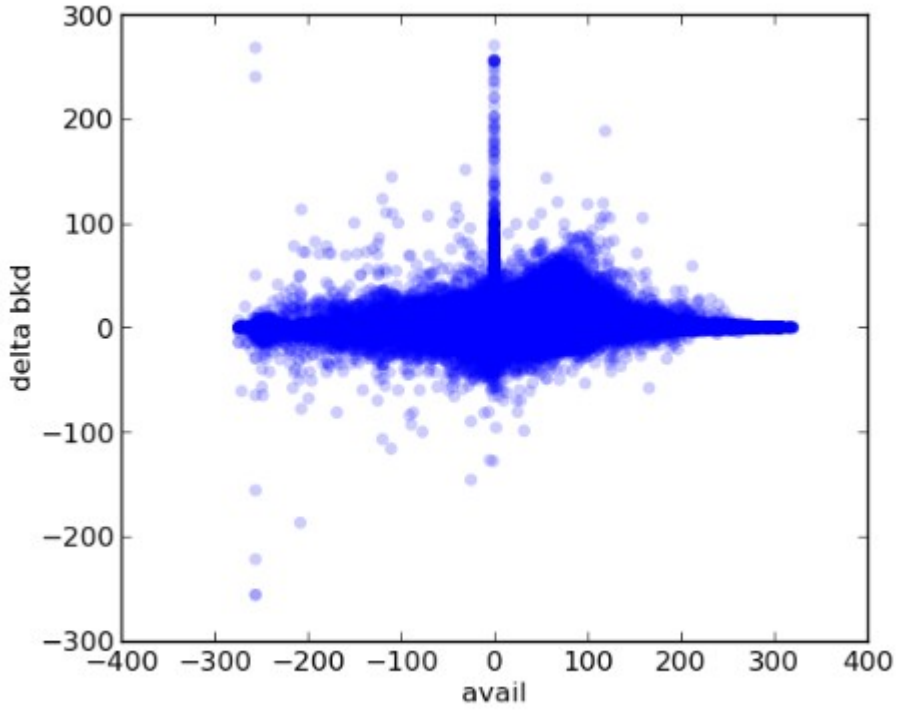


Figure 7: ΔBKD vs Keyday. As we approach departure, ΔBKD becomes more volatile.





Learning to Predict Bookings

Learning Framework

For our project we developed a model relating authorization curves to each flight's Cabin Load Factor (CLF) and the average cost of ticket. From the model we gained insight into the relative importance of different features in predicting cabin load factor and the distribution of ticket purchases across different booking classes.

Overview of Learners

Talk about why some learners are better than others...this is a 'question'.

Data Manipulation

Interpolation, removal of canceled flights, removing booking curves that didn't have any bookings, encoding of categorical variables

Feature Selection

Talk about how our data exploration informed our feature selection, reducing dimensionality of categorical features.

- bar chart of feature importances like categorical features for day of week (show that thursday, fri, saturday had higher importances than general weekdays. So we decided to reduce dimensionatity by compressing information into "weekend."

Model Performances

Discussion of how we assessed performance (K-fold cross validation, scoring, etc.), accuracy vs. percentage of data used

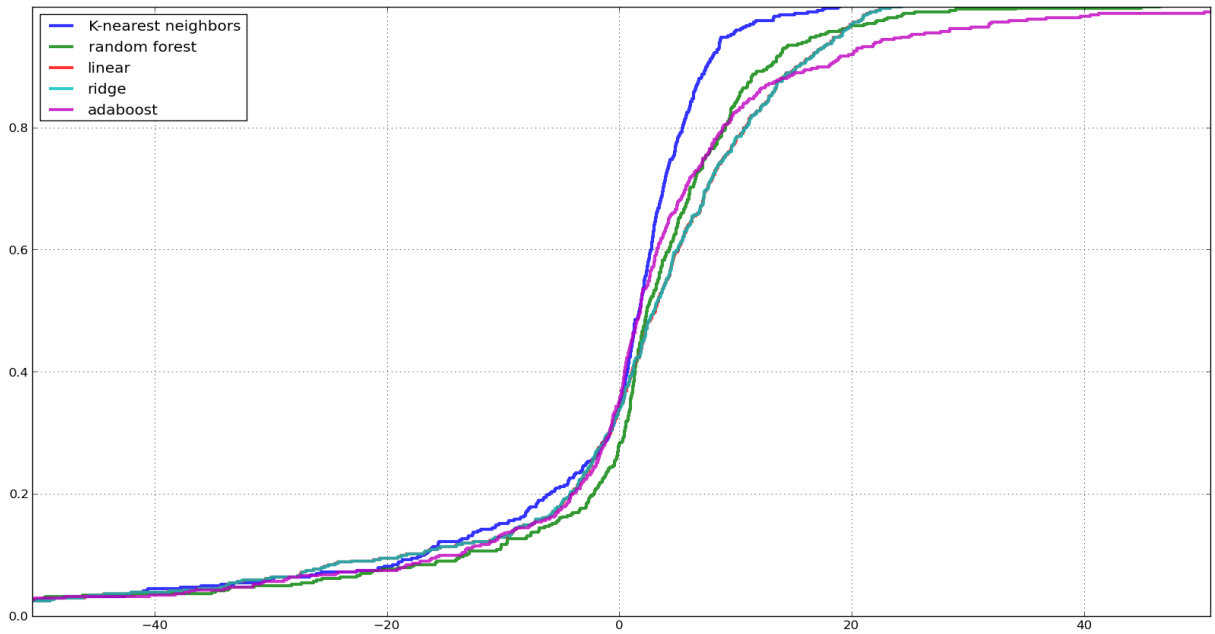
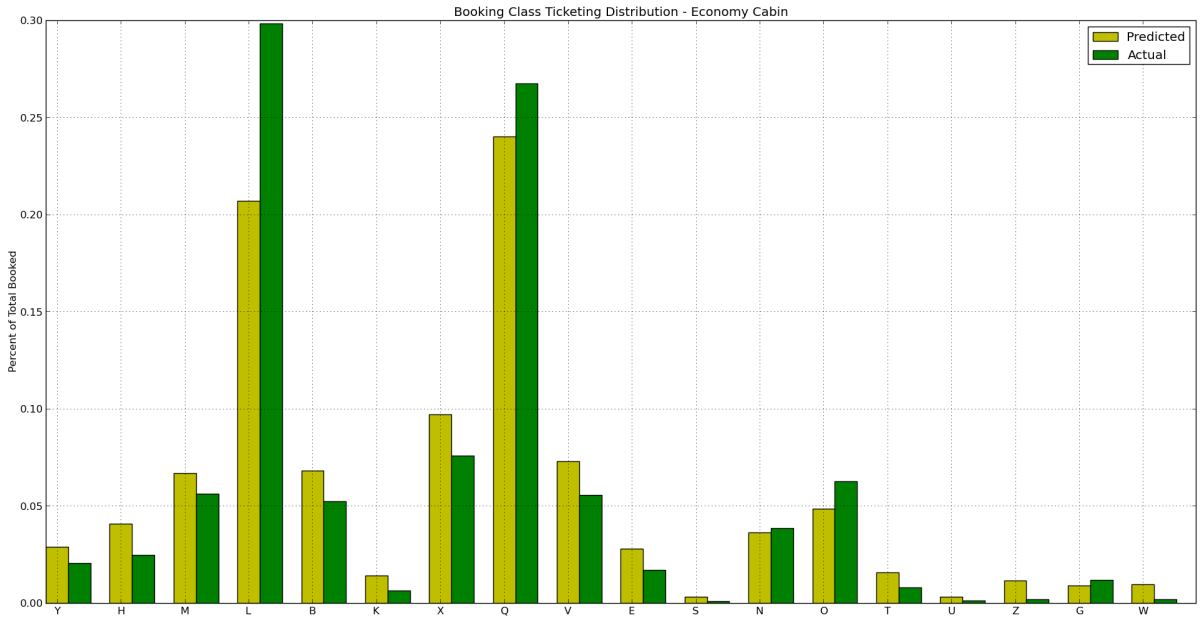


Figure 8: CDF of all errors in Total Booked on each flight. K-nearest neighbors performed the best; it never overestimated the number of people on a flight by more than twenty.

Analysis of KNeighborsRegressor



For best learner, graphs of 5-50-95 graphs and distribution across booking classes, representative booking curves

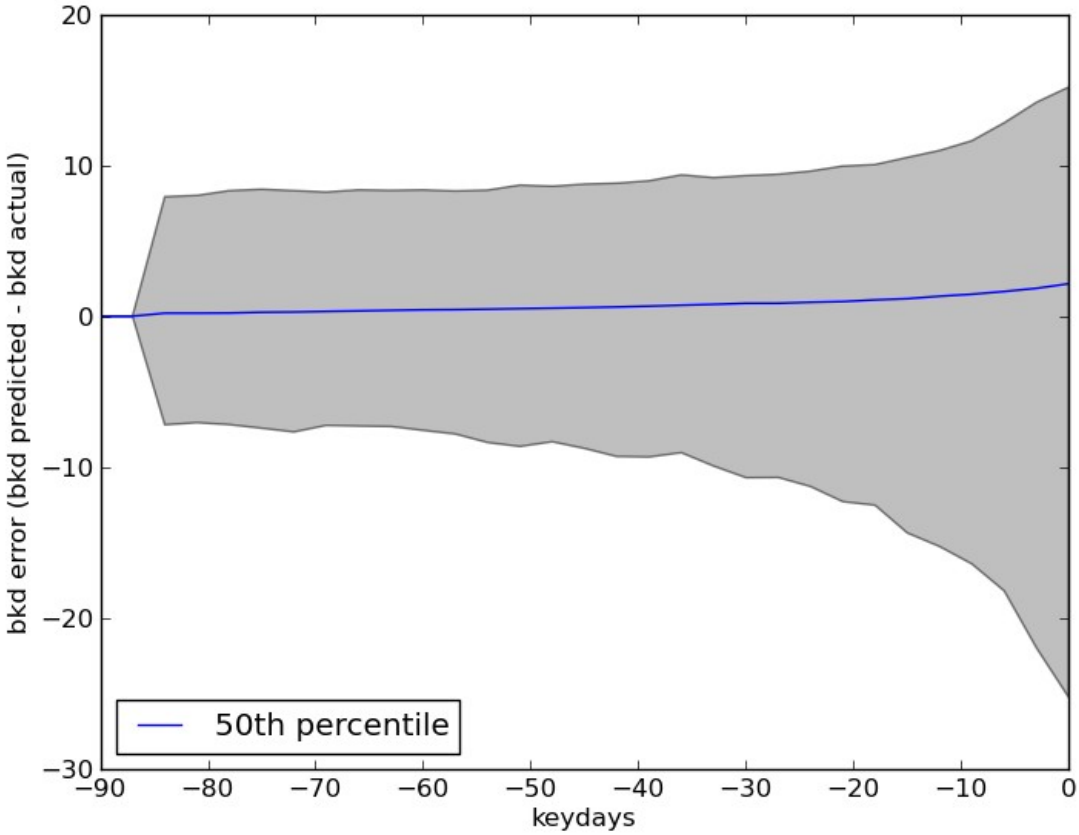
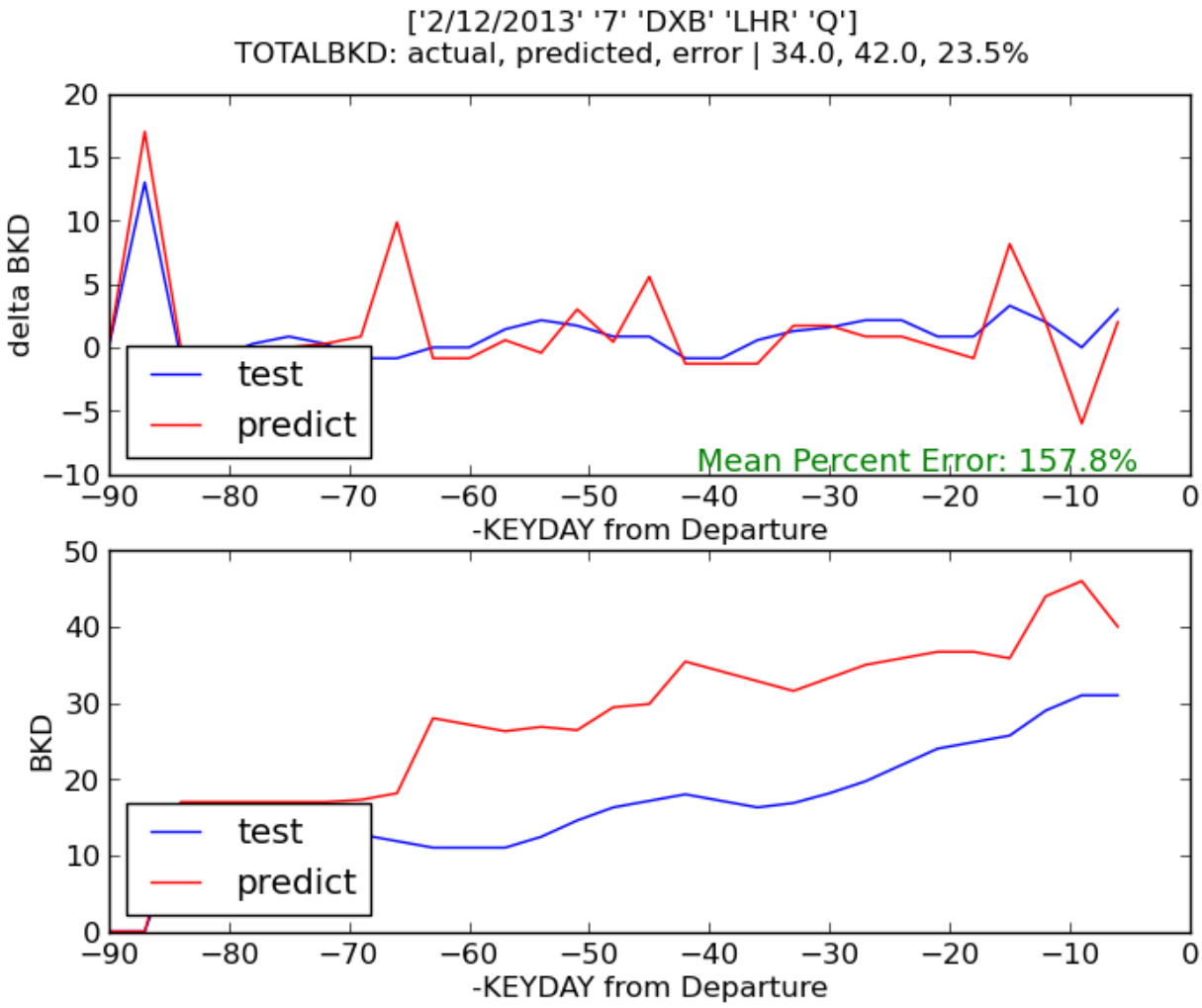


Figure 9: Cumulative error in total booked at each keyday. The gray region represents the inner 90 percent range of total booked error, at the 5 and 95 percentiles of the total booked CDF. The error starts to increase as the rate of bookings increases near departure.



Future Directions

Better learners...markov process controller...cool visualizations...larger data set...sequential forward feature selection...

Methodology

Predicting CLF is non-trivial because predicting TOTALBKD is non-trivial. Thus, we developed a model to predict TOTALBKD. TOTALBKD ican be calculated by the summation of deltaBKD over the entire ticketing period.
We decided to use a machine learning method to learn a model of our data. In particular,

Tried different learners...feature selection...cross validation...interpolation...removing data for canceled flights/booking curves that never get any bookings...categorical feature encoding/binning...feature extraction (like deltaBkdLower)...normalization?...Why feature selection? Include a graph of the model accuracy decreasing for additional features...Talk about how the inputs don't lend themselves to transforms that would make learning easier...talk about how we scored our model…

Results

+ Graphs of inputs vs. deltaBkd… aggregate
-Error CDFs of interpolated vs. not...
+graphs of error for different learners…
-tables of feature importances for best model…
-Graphs of Error for Different Feature Sets (specifically the different categorical encodings)...
-Histogram of Cabin Hierarchy vs Percentage of Cabin Filled

-GRAPHS OF OUTPUT FROM BEST LEARNER - GRAB A REPRESENTATIVE, INTERESTING BOOKING CURVE OR THREE, DISTIRBUTIONS OF BOOKING CLASS-NESS AGAINST EXPECTED

GRAPHS OF ACCURACY OF DIFFERENT LEARNERS - MAYBE SHOW HOW WELL THEY SCALE WITH MORE DATA? CDFS OF THEIR ERRORS (TOTALBKD, OVERLAYED ON THE SAME PLOT?