# An Agent for Optimizing Airline Ticket Purchasing (Extended Abstract)

William Groves and Maria Gini
Department of Computer Science and Engineering
University of Minnesota, USA
{groves,gini}@cs.umn.edu

## ABSTRACT

Buying airline tickets is an ubiquitous task in which it is difficult for humans to minimize cost due to insufficient information. Even with historical data available for inspection (a recent addition to some travel reservation websites), it is difficult to assess how purchase timing translates into changes in expected cost. To address this problem, we introduce an agent which is able to optimize purchase timing on behalf of customers. We provide results that demonstrate the method can perform much closer to the optimal purchase policy than existing decision theoretic approaches for this domain.

## Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Intelligent Agents*

## Keywords

airline ticket prices, feature selection, PLS regression

## INTRODUCTION

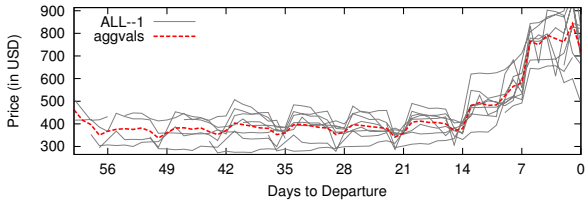
The conventional wisdom of airline ticket purchasing is to buy a ticket as early as possible to avoid the risk of price increase. However, as our analysis shows, the earliest purchase strategy only occasionally achieves the lowest cost. Often airlines lower prices for competitive reasons violating this conventional wisdom. Figure 1 shows systematic price patterns in an example of daily minimum prices for a route. This paper proposes a model for estimating the optimal buying policy for future departures. The ultimate application of this model is to autonomously make daily purchase decisions on behalf of airline ticket buyers to lower their costs.

Our approach relies on a corpus of historical data to compute policies that do much better than the earliest purchase strategy. The success of the proposed method depends on a user-generated hierarchical classification of the variables, which enables efficient discovery of a feature set with better performance over fully automated feature selection methods and avoids overfitting.

The approach is applicable to many real-world multivariate domains, but this paper demonstrates the power of this technique for airline ticket price prediction. An additional benefit is that the features selected provide insights into the domain: the importance of each variable can be assessed by its presence or absence in the computed optimal model.

**Appears in:** *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013)*, Ito, Jonker, Gini, and Shehory (eds.), May, 6–10, 2013, Saint Paul, Minnesota, USA.


**Figure 1: Lowest price from any airline for NYC→MSP 5-day round-trips departing on Thursday. A solid line indicates prices for a departure (8 departures in total). A dotted line is the mean.**




This paper was inspired by [1] where several purchasing agents attempt to predict the optimal purchase time (within the last 21 days prior to departure) of an airline ticket for a particular flight. The purchasing policy (a sequence of wait/buy signals) is computed for many unique simulated passengers with a specific target airline, target flight, and date of departure. Each of these simulated passengers represents a *purchase episode*, and the mean cost from many episodes is a measure of performance for a model (we use this measure for the experimental results reported here). We understand that Bing Travel's "Fare Predictor" is a commercialized version of the models in [1], and it provides another benchmark for our results.

Our work is different from [1] in that our model predicts the minimum cost ticket of *any* flight from *any* airline given a route and departure date. This is more difficult because the aggregate minimum price varies less than the price of an individual flight from an airline.

Fully automated feature selection methods are found in the literature including correlation-based and wrapper based techniques [2]. Our method improves performance by incorporating available domain knowledge in feature selection.

## PROPOSED MODEL AND RESULTS

In real world domains typically there are many variables (features) that may be relevant to the prediction, and feature selection is necessary to achieve good performance. We choose the best model as follows:

1. Feature Extraction – The raw data observed in the market are aggregated into a fixed length feature set. A detailed description of this process is available in [3].

**Table 1: Feature classes from general to specific**

| Class | # Vars | Variable List |
|---|---|---|
| D0 | 8 | days-to-departure, day-of-week |
| A1 | 3 | minimum price, mean price, and quote count for all airlines |
| A2 | 9 | statistics for non-stop, one-stop and two-stop flights |
| A3 | 18 | statistics for each airline |
| A4 | 54 | statistics for each airline and # stops |

**Table 2: Optimal lag scheme for a route**

New York → Minneapolis

| Class | Lagged Offsets | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| D0 | • | | | | | | | |
| A1 | • | • | • | • | • | • | • | • |
| A2 | | • | • | • | • | • | • | |
| A3 | | | | • | | | | |
| A4 | | | | | | | | |

The features are divided into *feature classes* (shown in Table 1) based on the strength of the relationship to the target feature: A1 is general, A4 is most specific.

2. Lagged Feature Computation – We introduce the lag scheme concept as a logical structure over the feature classes. A lag scheme defining a specific feature set (a subset of the standardized features) is computed using feature class constraints. *Lagged offsets* (time-delayed variable values 1..7) are also possible, but searching lags $> 7$ did not improve performance for this data set. More general classes must be included in the feature set before more specific classes (i.e. A1 must be included before A2). The constraints in this domain are A1 $\rightarrow$ A2 $\rightarrow$ A3 $\rightarrow$ A4. An example of a lag scheme that performs well for a target variable is shown in Table 2[1]. There are 8519 possible lag schemes over the 5 classes and time lags up to 8 days: exhaustive evaluation is feasible. An equivalent search over the 92 original features in the domain would be infeasible.
3. Regression Model Construction – Using the augmented feature set generated from the lag scheme, a regression model is generated using PLS regression or another machine learning algorithm as shown in the results.
4. Optimal Model Selection – Search all lag schemes and chose the best model based on a hold-out set.

Our experiments were designed to estimate real-world costs of using our prediction models. We assume, as airlines do, a relatively fixed rate of purchases until a flight is full, and that most tickets for a flight are sold within 60 days of departure [4]. There are two basic benchmarks used in testing: *earliest purchase* (each of its purchase episodes terminate with a purchase event on the first day of the episode and cost would be equal to the sum of prices observed over the 60 day period), and *optimal cost* (the sequence of buy/wait signals that leads to the lowest possible ticket price). These are used as benchmarks in the results. Bing Travel recommendations were also collected for the time period of the data set and its cost performance is shown. The objective is to achieve as close as possible to the optimal cost.

Table 3 shows the results of estimated costs for several purchasing policies based on 5-day round trip itineraries from NYC to MSP (265 simulated purchases per experiment). The results show how costs vary based on the model.

[1]D0, a feature class containing deterministic features, only appears at most once as lagged values from it can be deterministically computed from the most recent value.

**Table 3: Results comparison on a single route**

| Feature Selection | Learning Method | NYC-MSP Mon-Fri |
|---|---|---|
| (mean cost (in $), efficiency (as % of optimal savings)) | | |
| Benchmark | Earliest Purchase | (317, 0.00%) |
| | Optimal | (268, 100%) |
| | Bing Travel | (308, 2.56%) |
| No Feature Selection | PLS w/Minimal Lag Sch. | (314, 6.87%) |
| | PLS w/Full Lag Scheme | (300, 34.1%) |
| Off-the-shelf Methods | PLS w/Correlation-based FS | (313, 7.93%) |
| | PLS w/Wrapper FS (BFS) | (317, −1.21%) |
| Lag Scheme Feature Selection | Decision Tree | (288, 58.8%) |
| | nu-SVR | (295, 45.1%) |
| | Ridge Regression | (293, 49.9%) |
| | PLS Regression | (280, 75.3%) |

For comparison, experiments with no feature selection are also made to highlight the benefit of feature selection: Minimal Lag Scheme (provide only the most recent observation of each variable) and Full Lag Scheme (provide all lag offsets 0..7 for all variables). The lag scheme approach works well for many choices of machine learning algorithms, but PLS regression was found to work best for this domain. The improved performance can be attributed to a natural resistance to collinear and irrelevant variables even when there are relatively few observations. The lag scheme search constrains feature selection and avoids overfitting caused by little training data in large multivariate data sets.

# CONCLUSIONS AND FUTURE DIRECTIONS

To our knowledge, these results represent the state-of-the-art in airline ticket price prediction using consumer-accessible data. This investigation shows that, given sufficient publicly-observable information, it is possible to predict airline ticket prices sufficiently to reduce costs for customers. There is significant demand for deployed versions of this kind of prediction model in the hands of consumers.

Generating a feature set hierarchy requires some domain knowledge, but does not require expert level understanding. The inclusion of lagged features in the model captures temporal relationships among feature and improves the predictions. By examining the best lag schemes, domain knowledge can be extracted: the significance of individual features can be discovered by observing their presence in the scheme.

# REFERENCES


[1] Etzioni, O., Tuchinda, R., Knoblock, C., Yates, A.: To buy or not to buy: mining airfare data to minimize ticket purchase price. In: SIGKDD Conf. on Knowledge Discovery and Data Mining. (2003) 119–128

[2] Hall, M.A.: Correlation-based feature selection for discrete and numeric class machine learning. In Langley, P., ed.: ICML, Morgan Kaufmann (2000) 359–366

[3] Groves, W., Gini, M.: A regression model for predicting optimal purchase timing for airline tickets. Technical Report 11-025, University of Minnesota, Minneapolis, MN (2011)

[4] Belobaba, P.P.: Airline yield management. an overview of seat inventory control. Transportation Science **21**(2) (1987) 63–73