



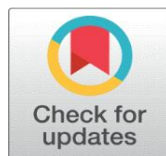


## PREDICTIVE MODEL FOR AIRLINES' FLIGHT DELAY & PRICING

Prashant Kapri <sup>1</sup>, Noopur Thanvi <sup>2</sup>, Shubham Patane <sup>3</sup>, Rashmi Thakur <sup>4</sup>

<sup>1</sup>Thakur College of Engineering & Technology Mumbai, India



### Corresponding Author

Prashant Kapri,  
[prashantkapri7@gmail.com](mailto:prashantkapri7@gmail.com)

### DOI

[10.29121/shodhkosh.v3.i1.2022.6022](https://doi.org/10.29121/shodhkosh.v3.i1.2022.6022)

**Funding:** This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

**Copyright:** © 2022 The Author(s). This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

With the license CC-BY, authors retain the copyright, allowing anyone to download, reuse, re-print, modify, distribute, and/or copy their contribution. The work must be properly attributed to its author.



### ABSTRACT

Now-a-days Airline ticket prices and delays in the flight have become unpredictable. Ticket prices are dynamic and change significantly for the same flight and even for the same class of seat. Airline companies implement various algorithms to change the prices dynamically, so as to maximize their revenue. Because of tough competition among airline services these models are not available to the general public. Also, the flight gets delayed because of various micro and macro factors. The major factors that affect the airlines are air route situation, delay of previous flight, aircraft capacity, air traffic control, airline properties, etc. There is a need to predict the flight delays and flight prices of airlines to save both 'Time and Money'. We are building a platform for airplane commuters to predict the flight delays and flight prices. Using this tool, they will be able to plan their travel efficiently and thereby save money. The interface of the tool will be user-friendly. We will be applying various machine learning algorithms to predict the prices and delays, and implement the most efficient and effective algorithms in the tool. Our system will comprise of two main components namely price prediction module and delay prediction module.

**Keywords:** Airline, Ticket Price, Delay, Machine Learning Algorithms

## 1. INTRODUCTION

Delay is one of the most remembered performance indicators of any transportation system. Notably, commercial aviation players understand delay as the period by which a flight is late or postponed. Thus, a delay may be represented by the difference between scheduled and real times of departure or arrival of a plane. Flight delays have negative impacts, mainly economic, for passengers, airlines, and airports. Given the uncertainty of their occurrence, passengers usually plan to travel many hours earlier for their appointments, increasing their trip costs, to ensure their arrival on time.

On the other hand, airlines suffer penalties, fines and additional operation costs, such as crew and aircrafts retentions in airports. Furthermore, from the sustainability point of view, delays may also cause environmental damage by increasing fuel consumption and gas emissions. Subsequently, airline ticket prices can vary dynamically and significantly for the same flight, even for nearby seats within the same cabin. Customers are seeking to get the lowest price while airlines are trying to keep their overall revenue as high as possible and maximize their profit. Airlines use various kinds of computational techniques to increase their revenue such as demand prediction and price discrimination. From the customer side, two kinds of models are proposed by different researchers to save money for customers: models that predict the optimal time to buy a ticket and models that predict the minimum ticket price.

This report seeks to summarize the most researched trends in this field, describing how this problem is addressed and comparing methods that have been used to build prediction models. This becomes more relevant as we observe an increasing presence of machine learning methods to model flight delays predictions. The last two decades have seen steadily increasing research targeting both customers and airlines. Customer side researches focus on saving money for the customer while airline side studies are aimed at increasing the revenue of the airlines. Conducted researches employ a variety of techniques ranging from statistical techniques such as regression to different kinds of advanced data mining techniques.

## 2. LITERATURE REVIEW

One of the pioneers on ideal ticket buy timing expectation is most likely the work done by (Ethion et al, 2003). The creators proposed a model that exhort the client whether to purchase a ticket or to hold up at a specific purpose of time. For each question day, the model creates a purchase or hold up signal dependent on recorded value data. The model uses different information mining systems, for example, Rule learning (Ripper), Reinforcement learning (Q-learning), time arrangement techniques, and blends of these to accomplish different precision levels. Q-learning and Ripper are utilized to anticipate the conduct of new flight information dependent on a lot of preparing information while the time arrangement strategy utilizes the moving normal to conjecture the value attributes of a flight dependent on authentic value information of a similar flight. Around 12,000 chronicled ticket value information speaking to 41 flight dates for two courses was utilized for the examination.

The dataset has restrictions in which the assortment was done just beginning from twenty-one days before the flight. Additionally, a consistent seven days full circle is considered. The highlights utilized by the model incorporate flight number, number of hours until takeoff, current value, carrier and course (root and goal city). Recreation is utilized to gauge the investment funds travelers increased because of every one of these information mining techniques. The sparing (or misfortune) execution of a model is determined by figuring the expense because of the value contrast between the ticket cost at a prior buy point and the ticket cost at the time suggested by the calculation. The best precision (61.9% when contrasted with ideal sparing) is accomplished from the blend of the considerable number of procedures utilized in the investigation. As indicated by (William Groves [4]), the model proposed by (Etzioni [6]) possesses been executed in genuine energy for a mainstream ticket search site referred to as Bing Travel as "Passage Predictor" instrument.

A firmly related work to that of (Etzioni [6]) is likewise proposed by (William Groves [4]) which predicts ideal ticket buy timing and is in reality propelled by (Etzioni [6]). Be that as it may, not at all like (Etzioni [6]), (William Groves [4]) can gauge the ideal buy time for every single accessible trip crosswise over various carriers for a given takeoff date and course. In addition, they utilize a dataset that is gathered 60 days in front of the flight date.

The information was gathered for a time of 3 months utilizing every day value cites from an OTA site from February 22, 2011 to June 23, 2011. Each question returned roughly 1,200 statements for a solitary course from all aircrafts. The full circle depended on a consistent 5 days full circle. Two sorts of highlights were utilized for the investigation: Deterministic includes and collected highlights. Instances of deterministic highlights incorporate days to flight and cite day of week for example the quantity of various ticket costs accessible for a given trip on a particular course crosswise over various carriers). Collected highlights are highlights removed from the verifiable information, for example, the base value, mean cost and number of statements.

The base value, mean cost and number of statements are determined for constant, one-stop and multi-stop for singular aircrafts and for all carriers. In addition, a slacked highlight calculation is likewise used to consider the impact of time-deferred perceptions in forecast. Four sorts of relapse methods are utilized to create a relapse model for the examination: Partial least squares (PLS) relapse and three AI calculations (Decision tree, nu-Support Vector Regression (nu-SVR) and Ridge Regression).

The examination in (William Groves [4]) utilized a comparative methodology as (Etzioni [6]) for the presentation assessment. In light of trial examination for one course with 256 recreated buys, PLS relapse was seen as the best model with 75.3% sparing when contrasted with the ideal one.

**Table 1**

Tabular Representation of the Various Literature Survey

Year	Author	Main Objective	Methodology	Limitations
2019	Juhar Ahmed Abdella , Nazar Zaki[2]	The automatic creation of literature abstracts	1.Ignore stop words 2.Determine frequently occurring/top words 3. Select top words 4. Select top sentences	Does not consider heterogenous flights.
2018	Bin Yu , Zhen Guo , Sobhan Asian[1]	Flight Delay Prediction	DBN-SVM	Arrival and international flights information were not used to predict delays. Cargo flights were eliminated in the delay prediction because the “planned waiting” data of cargo flights were not available.
2013	William Groves and Maria Gini[4]	Predicting optimal ticket purchase time	PLS regression Decision tree, nu – SVR Ridge Regression	Does not consider heterogeneous flights.
2013	Arikan et al.[3]	Delay Prediction	Queuing Model	In the source document is quite small (about 1 paragraph or ~500 words in the training dataset)
2014	Rebollo and Balakrishnan [5]	Delay Prediction	Random Forest	Focuses on network (route) delay rather than individual flight.  Does not capture localized delays (for example, mechanical issues)
2003	Etzioni et al.,[6]	Predicting optimal ticket purchase time	Rule learning (Ripper), Reinforcement learning (Q-learning), time series methods.	Multi-leg flights not included.

### 3. DATA COLLECTION

The assortment of information is the most significant part of this venture. There are different attributes of the information on various sites which are utilized to prepare the models. Sites give data about the different routes, times, carriers, fare, delays and much more. Different sources from government websites to buyer travel sites and airline portals are considered for information scratching.

#### 1) Overview of Dataset

The dataset has been taken from a online machine hack competition [..]. The size of training set is 10683 data records and testing set is of size 2671 records. It consists of various attributes which are as follow.

**Table 2**

DESCRIPTION OF ATTRIBUTES INVOLVED IN THE PRICING DATASET

<i>Attribute</i>	<i>Description of Attributes</i>
Airline	The name of the airline.
Date_of_Journey	The date of the journey
Source	The source from which the service begins.
Destination	The destination where the service ends.
Route	The route taken by the flight to reach the destination.
Dep_Time	The time when the journey starts from the source.
Arrival_Time	Time of arrival at the destination.
Duration	Total duration of the flight.
Total_Stops	Total stops between the source and destination.
Additional_Info	Additional information about the flight

Price	The price of the ticket
-------	-------------------------

The dataset has been taken from a reliable online available government agency website that provides the air traffic delay statistics in the United States. The Dataset name is Reporting Carrier On-Time Performance (1987-present). Data is provided by Bureau of Transportation Statistics of United States. Data is downloadable in Monthly Format. The data that we have considered is from July 2017 – June 2019 [2 Years]. It consists of various attributes which are as follows.

**Table 3**

**DESCRIPTION OF ATTRIBUTES INVOLVED IN THE DELAY PREDICTION DATASET**

<i>Attribute</i>	<i>Description of Attributes</i>
Reporting Airline	The name of the airline.
Flight Date	The date of flight broken down in day, month, quarter and year.
Tail Number	Identification code of unique aircraft
Origin	Origin city/town and also includes state.
Destination	Destination city/town and also includes state.
CRS_Dep_Time	Expected departure time.
CRS_Arrival_Time	Expected Arrival time.
Arr Time	Arrival Time.
Arr Delay	Arrival Delay
ArrDelay15	Arrival Delay with class 0(<15) and 1(>15)
Distance	Distance between airports.

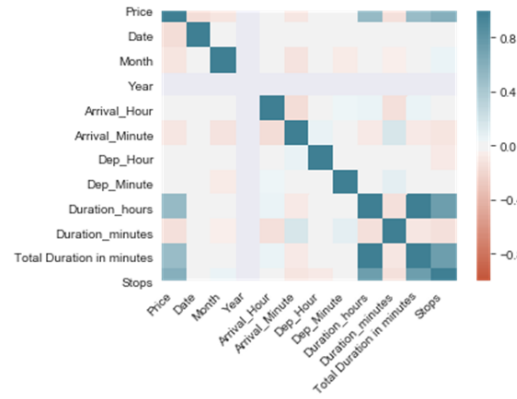
- Data Preparation**

The kind of data that we downloaded was very raw and needed a lot of work. For instance, the Duration was a character type and not an integer. Moreover, for any model to work efficiently, certain variables need to be introduced by combining or changing the existing variables. We selected the features and found their correlation also performed various visualisation techniques on the data. After removal of null values the data retained was 99%. We also performed label Encoding on columns such as Additional\_Info, Airline, Destination, Source, Route\_1, Route\_2, Route\_3, Route\_4, Route\_5

For Delay prediction data we gathered the data from 'Transat' website. Data was in monthly format so we merged the data. We removed Null Values and found the data retention to be 98%. We carried out test on data like implementing correlation to find relationships among features. We performed Label Encoding on "Reporting Airlines", "Tail Number", "Origin", "Origin State", "Destination" and "Destination State". We did this to convert categorical values into numerical values.

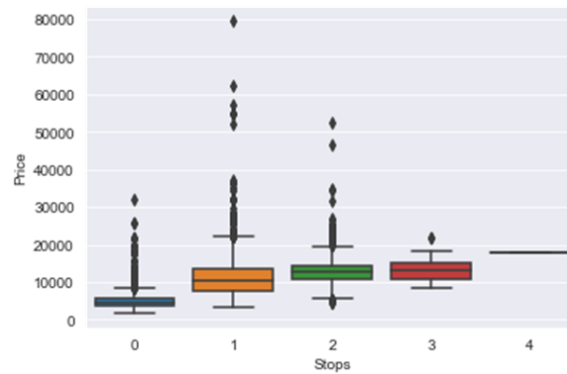
- Data Exploration and Analysis**

Furthermore, data exploration and data analysis take place to uncover and reveal the hidden trends in the data. This analysis helped us in understanding the data and its feature more clearly, also it made clear to us about the features which we can use in the machine learning model.



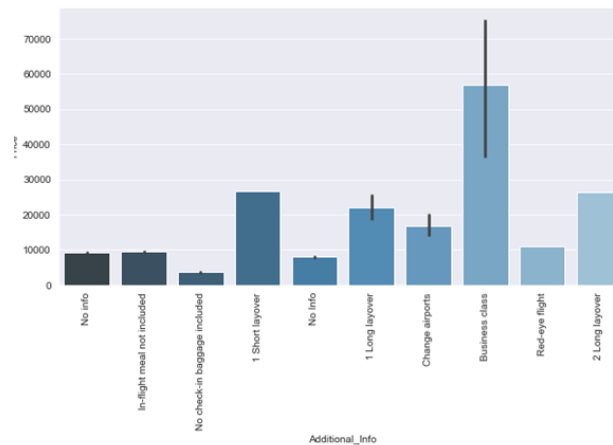
**Figure 1** HEATMAP OF PRICING DATASET

The above figure visualizes the correlation between Price attribute with all the attributes present in the data. One can easily figure out by looking that Price attribute has a strong correlation with the Duration of flight and the no of stops a particular flight have.



**Figure 2** VARITIONS IN PRICE W.R.T NO. OF STOPS

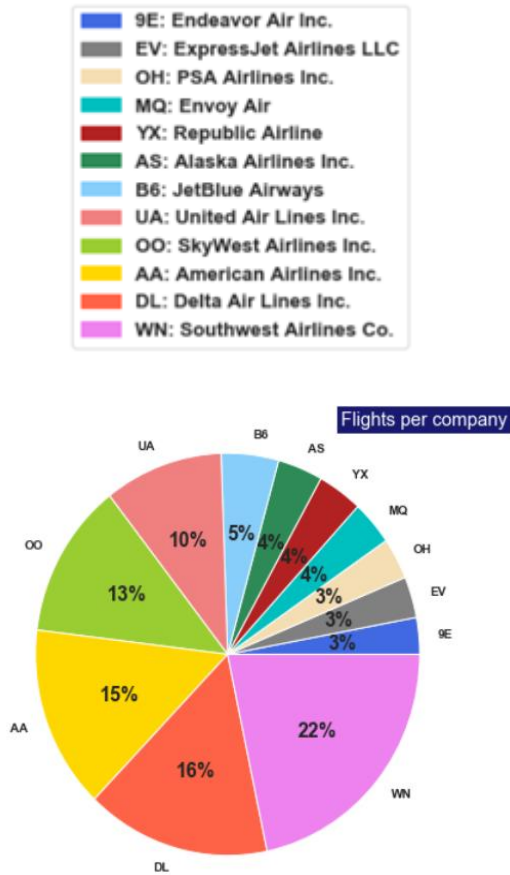
The above figure helps us in knowing the statistical value of the data such as mean, median, mode and the most important it identifies and visualize the various outliers present in the data.



**Figure 3** TRENDS IN PRICE W.R.T ADDITIONAL FACILITES PROVIDED IN THE FLIGHT

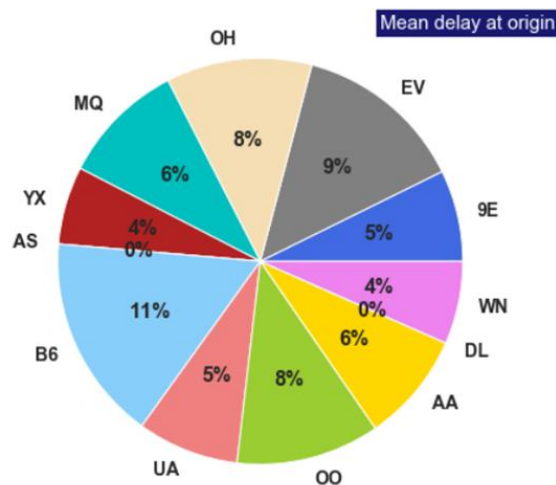
The above graph visualizes the trends in airplane prices with the additional facilities provided during the journey. Such as journey in the business class has the highest price.

### Visualizations:



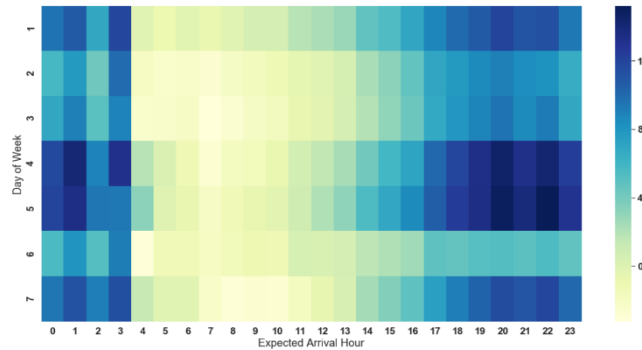
**Figure 4** SHARE PERCENTAGE PER AIRLINES

We can observe in Fig1 that there are 14 domestic airlines operational in US. Out of 14 only 4 constitutes to 75% of flights. Southwest, Delta airlines and American airlines have largest volume of domestic flights.



**Figure 5** MEAN DELAY AT ORIGIN PER AIRLINE

From this Pie chart we can clearly note that the mean delay at origin is almost evenly distributed. Even though in above pie chart we found that some airlines are more in count, we observe that it dose not affect the mean delay.



**Figure 6** HEATMAP OF DAY OF WEEK Vs EXPECTED ARRIVAL TIME

Heatmap gives a clear representation of at what time of day the delay is most probable. It can be conclude from the heatmap that from 10 PM to 3 AM delay is the most. Also from 4 AM to 1 PM the delay is the least. As compared to other days Thursday and Friday has the most delay

#### 4. MACHINE LEARNING ALGORITHM

To build up the model for the flight price prediction, numerous regular machine learning algorithm were applied. They are as per the following: Linear regression, Ridge regression, Lasso Regression, XGBoost, Light GBM, and Stacking model. Each one of these models are executed in the scikit learn. To assess the presentation of this model, certain parameters are considered i.e. Root Mean Squared Error (RMSE). The equation for the above parameters is as follow:

Root Mean Squared Error (RMSE) is the square root of the mean of the squared errors:

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}}$$

##### 1) Linear Regression

Regression is a method of modeling a target value based on predictors that are independent. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. A linear regression line has an equation of the form

$$Y = a + bX,$$

where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when x = 0).

##### 2) Ridge and Lasso Regression

Regularization is a method which makes slight alterations to the learning method with the end goal that the model sums up better. This indirectly improves the model's performance on the hidden information too. Basically, it is used to reduce the overfitting in the learning process and deal with the outliers present in the dataset. In Regularization process the tuning of function takes place by addition of a penalty term in the error function. The additional term controls the excessively fluctuating function such that the coefficients don't take extreme values.

- **L1 Regularization aka Lasso Regularization:** This add regularization terms in the model which are function of absolute value of the coefficients of parameters. The coefficient of the paratmeters can be driven to zero as well during the regularization process. Hence this technique can be used for feature selection and generating more parsimonious model



- **L2 Regularization aka Ridge Regularization:** This add regularization terms in the model which are function of square of coefficients of parameters. Coefficient of parameters can approach to zero but never become zero and hence

### 3) XGBoost

XGBoost is an ensemble learning method. Sometimes, it may not be sufficient to rely upon the results of just one machine learning model. Ensemble learning offers a systematic solution to combine the predictive power of multiple learners. The resultant is a single model which gives the aggregated output from several models. XGBoost (extreme Gradient Boosting) is an advanced implementation of the gradient boosting algorithm. XGBoost has proved to be a highly effective ML algorithm, extensively used in machine learning competitions and hackathons. XGBoost has high predictive power and is almost 10 times faster than the other gradient boosting techniques. It also includes a variety of regularization which reduces overfitting and improves overall performance.

### 4) Light GBM

Light GBM is a gradient boosting framework that uses tree-based learning algorithm. Light GBM grows tree vertically while other algorithm grows trees horizontally meaning that Light GBM grows tree leaf-wise while other algorithm grows level-wise. Light GBM is prefixed as 'Light' because of its high speed. Light GBM can handle the large size of data and takes lower memory to run. Another reason of why Light GBM is popular is because it focuses on accuracy of results. LGBM also supports GPU learning.

### 5) Random Forest

It is an ensemble learning method for regression and classification. The advantage of random forest is that it can be used for both relapse and characterization problem which most recent Machine learning algorithms have. Finding the significance of each feature is done easily in Random Forest. It forms many decision trees and adds them to get a stable expectation.

### 6) K-Nearest Neighbours

It is very robust for example classes do not have to be separable linearly. Irrelevant attributes

### 7) Stacking Model

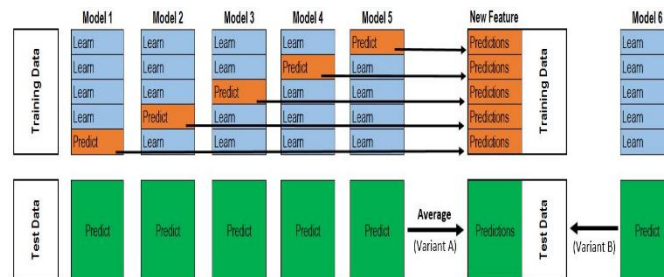


Figure 7 STACKING CONCEPT

We have used the concept of stacking for price prediction. Stacking means an assembling of all the models implemented into a single learning model as displayed in the above image.

In Price prediction we implemented various machine learning regression algorithm such as linear regression, Ridge regression, Lasso regression, Elastic net regression, XGBoost, Light GBM regression and much more and the calculated the Root Mean Square Error (RMSE) value for it and later compared it with the RMSE value of the model obtained from the stacking part.

## 5. RESULT AND CONCLUSION

Table 4 ALGORITHM EVALUATION FOR PRICING DATASET

Regression algorithms	RMSE value
Linear Regression	3237.8
Ridge Regression	3237.6
Lasso Regression	3237.8



Elastic net Regression	3237.8
XGBoost	1331.2
Light GBM	1428.2
Stacking model	1454.7

Thus, from the above figure we can conclude that the XGBoost and Stacking model gives the best accuracy as both the model has lowest root mean square error value.

**Table 5** ALGORITHM EVALUATION FOR DELAY DATASET

Classifier	Model Accuracy
Decision Tree	78
Naive Bayes	80
Neural Network	81
KNN	79
Random Forest	81
Linear Discriminant Analysis	81

As we can observe from the above table that the actual accuracy will always be high since the dataset is unbalanced. 80% of the flights are not delayed.

The main goal was to predict the remaining 19% data accurately. Almost all of the models have almost the same delay accuracy but Random Forest (Holidays and weather included) have the highest accuracy and precision in predicting delays.

**Table 5** ALGORITHM EVALUATION FOR DELAY DATASET [USING BALANCED DATA]

Classifier	Model Accuracy
Decision Tree	60
Gradient Boosting Classifier	62
XGB Classifier	62
KNN	58
Random Forest	64
Linear Discriminant Analysis	59

Here we have balanced the data. For balancing we took all the records having delays in them. Then we sampled equal number of records from the data which dose not have delay. For this operation we used "ArrDel15" field. Then we merged and shuffled both of them. After applying above algorithms we found the given accuracies. We can infer from these accuracies that Random Forest algorithm gives us the best results at predicting whether the flight will get delayed or not. It gave an accuracy of 64% which may not be good but is decent enough, given the data and how much unpredictable it is to predict delays.

## CONFLICT OF INTERESTS

None.

## ACKNOWLEDGMENTS

None.

## REFERENCES

- G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," *Phil. Trans. Roy. Soc. London*, vol. A247, pp. 529–551, April 1955. (references)
- J. Clerk Maxwell, *A Treatise on Electricity and Magnetism*, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in *Magnetism*, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

K. Elissa, "Title of paper if known," unpublished.

R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.