# Predicting Airline Delays: A Comparison of Models and Features

Dorian Goldman[1]

[1] *Conde Nast, Data Scientist,*
*Columbia University, Adjunct Professor of Data Science*

In this note, we compare different methodologies for predicting airline delays. We focus on all airline delays which are made publicly available on the website of The Bureau of Transportation Statistics []. We compare different modeling methodologies including classification and regression before and after an enriched set of features is introduced. The features which we add to provide supporting predictive accuracy are weather data from the NOAA, plane data from TODO and publicly available holiday data. We find that all three of the above improve predictive performance for regression and classification. However by implementing an auto regressive time series model to predict the mean delay on days prior to the day being evaluated, we find the strongest increase in performance. A more logical and robust approach would be to model the delays as a Poisson regression with the time dependent mean being a linear function of the features above, given the observed distributions of the data. However given the time constraints, this approach was not investigated further. This observation confirms the paradigm that carefully constructed features are often more predictive than more complex models in general.

## I. INTRODUCTION

Airline delays are becoming an increasingly problematic issue for many people around the world. With increased competition amongst airlines [] and increasingly restrictive choices for customers [] along with limited resources [], the problem has been further amplified. Flights which are delayed or canceled have cost consumers millions of dollars over the past few years alone [], and this trend is only seeming to increase []. To this date, there has been little effort to provide reasonably accurate predictions about delays which could potentialy save consumers a great deal of money and headache. While there have been modest attempts to understand this phenomenon [], most approaches fall short of a robust method of modeling these delays that can be used by the consumer in advance. In this paper, we aim to approach the problem from the approach of machine learning and to compare the power of various models to that of careful feature selection.

## II. DISCUSSION OF MAIN RESULTS

### A. Regression on Departure Delay

Use `sections` and `subsections` to organize your document. LaTeX handles all the formatting and numbering automatically. Use `ref` and `label` for cross-references — this is Section **??**, for example.

### B. Classification on Departure Delay

### C. Most prevalent features

Use `tabular` for basic tables — see Table I, for example. You can upload a figure (JPEG, PNG or PDF) using the files menu. To include it in your document, use the `includegraphics` command (see the comment below in the source code).

| Item | Quantity |
|---|---|
| Widgets | 42 |
| Gadgets | 13 |

TABLE I. An example table.

### D. Conclusion

LaTeX is great at typesetting mathematics. Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$, and let

$$S_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i}^{n} X_i$$

denote their mean. Then as $n$ approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$.

### E. Lists

You can make lists with automatic numbering ...

1. Like this,
2. and like this.

... or bullet points ...

- Like this,
- and like this.

**ACKNOWLEDGMENTS**

We thank...