# Flight Delay Prediction

**Sirjan Kafle**
sxkafle12@gmail.com
https://flightdelay.us

## 1 Introduction

We build an app that models the distribution of flight delays conditioned on variables like:

- Origin Airport
- Destination Airport
- Airline
- Time of Day

More specifically, we want users to provide a subset of those "conditional variables" and we'll output information on the probability that the delay time will exceed certain thresholds, e.g. 30 minutes, 1 hour, etc along with plotting those probabilities vs delay time. See figure (TODO attach plot) as an example.

We are resource constrained based on the container size of the web app we're deploying to, so we describe compression methods for these distributions by modeling them as exponentials - we store and load the parameters of that distribution for the app rather than full flight data.

In this doc, we explain a general framework for building algorithms to efficiently model these probabilities at scale. We start with describing how we model conditional distributions as an exponential, then describe an algorithm to handle low data support conditionals - e.g. a configuration of Origin Airport, Destination Airport, Airline, Time of Day that has very few data points that we can't give a reasonable estimate directly. In these cases, our algorithm is a Bayesian model with some assumptions. By doing so, our app can now handle queries for any subset of the conditional variables. We denote cases where we have low data support in the app marked by an asterick, see figure (TODO attach plot).

## 2 Modeling Conditional Distributions as an Exponential

### 2.1 Motivating Example

TODO: show plots - full and conditioned on $D > 0$.

### 2.2 Proposed Model

$$f_D(\delta|\{C_i\}) = p\lambda e^{-\lambda\delta}$$

### 2.3 Maximum Likelihood Estimate of Parameters

Learn $p$, $\lambda$.

$$p = \sum_{i=1}^{n} 1_{\delta_i > 0}/N$$

$$\lambda = N/\sum_{i=1}^{n} \delta_i 1_{\delta_i > 0}$$

### 2.4 MLE Plot and Goodness of Fit

#### 2.4.1 Metric

#### 2.4.2 Example Fits

#### 2.4.3 Score vs Data Size

#### 2.4.4 Score vs Conditional Group Profile

#### 2.4.5 Bias Variance Tradeoff in Action

### 2.5 Delay Survival Function

$$P(D > \delta | \{C_i\}) = pe^{-\lambda\delta}$$

## 3 Algorithms for Low Data Support Conditionals

### 3.1 Core Problem and Motivating Example

### 3.2 Definitions of Subset Groupings and Partition

### 3.3 Assumption and Setup

### 3.4 Parameter Estimate Derivation Given Partition

### 3.5 Which Partition is Best?

#### 3.5.1 Main Idea

#### 3.5.2 KL Divergence Derivation

#### 3.5.3 Ranking of Partitions Algorithm

## 4 Final Algorithm

### 4.1 Model Fitting

Defer explanation of the partitions algorithm to earlier.

### 4.2 Inference

## 5 Fun Examples

Attach images from the site