Predicting Flight Delays with Deep Learning
Ethan Chi, Jillian Tang
{ethanchi, jilingt}@stanford.edu

**Problem**
Given a flight and associated weather data, classify it as delayed (>15 min) or non-delayed.

**Background**
- With 24% of flights in the US delayed, delays are extremely costly both to consumers and airlines.
- Due to the sequential nature of flight delays, previous machine learning approaches have been mostly unsuccessful [9].
- Sequential DL models can better predict delays.

**Data**
- 314,677 flights departing from Chicago O’Hare International Airport in 2009 (split 80%-10%-10%)
- Time-matched weather data from the NOAA
- Total: 325 features, including:
  - **Weather**: temperature, dew point, cloud coverage, snow, rain, wind speed, wind direction, etc.
  - **Categorical**: carrier, airport (arrival + destination)
  - **Other**: Time, date, arrival delay (= ground truth)

**Baseline Models**
- **GBMs**
- **Logistic Regression**
- **Naïve Bayes**

**Results**

<table>
<thead>
<tr>
<th>Model</th>
<th>GBM</th>
<th>Log. Reg.</th>
<th>Naïve Bayes</th>
<th>LSTM</th>
<th>CNN+LSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>81.11%</td>
<td>81.02%</td>
<td>44.49%</td>
<td>78.89%</td>
<td><strong>81.90%</strong></td>
</tr>
<tr>
<td>Precision</td>
<td><strong>75.08%</strong></td>
<td>64.71%</td>
<td>23.95%</td>
<td>44.03%</td>
<td>52.32%</td>
</tr>
<tr>
<td>Recall</td>
<td>12.50%</td>
<td>17.39%</td>
<td><strong>77.47%</strong></td>
<td>43.93%</td>
<td>46.52%</td>
</tr>
<tr>
<td>F1</td>
<td>21.43%</td>
<td>27.41%</td>
<td>36.64%</td>
<td>43.98%</td>
<td><strong>49.25%</strong></td>
</tr>
</tbody>
</table>

**Methods**

**Model 1**: CNN + LSTM (based on LSTMNet [20])
- Previous flight data fed through CNN into LSTM
- Output combined with weather data, fed into dense layers

**Model 2**: same as above, but with a dense neural network instead of a CNN [20].
We trained 500 epochs with tf.train.AdamOptimizer.
Batch size = 5000. L1 regularization applied with λ=0.005
Class imbalance: weighted \( L = -2.6y \log \hat{y} - (1 - y) \log(1 - \hat{y}) \)

**Discussion**
- As expected, our model greatly outperformed baseline statistical models.
- We were able to replicate the performance of the state-of-the-art paper in the field [20].
- Using CNNs for preprocessing led to better performance, and due to the cheaper nature of convolutions, they were much faster than preprocessing with several dense layers.

**Future Work**
- Inclusion of data from more airports and more years to increase training set diversity
- More complex architectures (e.g. multiple layers of LSTMs)
- Integration of more features (national threat level, economy)

**References**