FLIGHT DELAY PREDICTION

PROJECT PRESENTATION

- **DATASETS:** Our main dataset includes every domestic flight from 2001 to 2008, with scheduled times, actual times, airports (departure/arrival), delay (if any) and cause of delay, plane identifier.

- **GOAL:** To predict if a flight will be delayed at arrival or not.
  - Hard to train a model to capture the delay of previous flights.
  - Imbalanced dataset between delayed and non-delayed examples.

EXPLORATORY DATA ANALYSIS

- **FEATURE ENGINEERING**
  - Description of some features added:
    - **attack_911:** If the flight took place on September 11th.
    - **attack_open:** If the flight took place in the post-9/11 period.
    - **airline:** The airline of the flight.
    - **weather:** Weather conditions on the day of the flight.
    - **airports:** Closest airports to the origin/destination.
    - **delay_reason:** Reason for the delay.
    - **aircraft:** Aircraft type.
    - **month:** Month of the flight.
    - **dayofyear:** Day of the year.
    - **date:** Date of the flight.
    - **time:** Time of the flight.

- **FEATURE ANALYSIS**
  - Weights for the logistic regression on normalized features.
  - Several features of force memory are present.

ALGORITHMS

- **RANDOM FOREST**
  - Formula: $p(y|x) = \frac{1}{Z} \sum_{x \in T(x)} p(x) p(y|x)$

- **GAUSSIAN NAIVE BAYES**
  - Model: $p(x_1, \ldots, x_n | y) = \prod_{i=1}^n p(x_i | y)$
  - Training: $\phi_{y=1} = \sum_{y=1}^n \frac{1}{\sum_{y=1}^n 1[y^{(i)} = 1]} \phi_{y=1} = \sum_{y=1}^n \frac{1[y^{(i)} = 1]}{n}$

- **LOGISTIC REGRESSION**
  - Model: $p(y = 1 | x ; \theta) = \frac{1}{1 + e^{-\theta^T x}}$
  - Training: $\theta_j = \frac{1}{m} \sum_{i=1}^m \left( y^{(i)} - \hat{y}_n(x^{(i)}) \right) x_j^{(i)}$

- **NEURAL NETWORK**

RESULTS

<table>
<thead>
<tr>
<th>ALGORITHM</th>
<th>PRECISION</th>
<th>RECALL</th>
<th>F1-SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>77%</td>
<td>100%</td>
<td>87%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>86%</td>
<td>91%</td>
<td>89%</td>
</tr>
<tr>
<td>Gaussian Naive Bayes</td>
<td>82%</td>
<td>88%</td>
<td>85%</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>74%</td>
<td>76%</td>
<td>75%</td>
</tr>
<tr>
<td>Neural Network</td>
<td>93%</td>
<td>73%</td>
<td>82%</td>
</tr>
</tbody>
</table>

OPTIMIZATION OF THE RANDOM FOREST

- **F1-score vs max_depth**
- **F1-score vs min_samples**

ANALYSIS

- **NEXT STEPS:**
  - Improve the Neural Network (better shape).
  - Increase size of the sample of training examples with data from other years.
  - Add new features to capture non-linearity (we still have no overfitting).