### Introduction

This part of project focused on applying Support Vector Machine (SVM) classifier to predict flight delays. Priority was made to maximize the Recall metric for delayed flights. It is because it is better if the model incorrectly predicts a flight as delayed than if it fails to classify it, as it could lead to disruption in many works at the airport such as: gate assignments, crew scheduling, baggage handling or passenger connections.

It was observed throughout the testing that a strong trade-off exists between recall and precision, meaning good recall generally resulted in poorer precision

### Environment and setup

The RAPIDS-Colab was used to train GPU, along with libraries such as cuml, cupy, sklearn, and optuna for execution and optimization

The preprocessed flight data was loaded from a file that contained shuffled 100 k rows from original dataset, which was made to decrease processing time. For training and test first 50 000 rows were used.

The target variable for prediction is is\_arr\_delayed. Categorical features were encoded by LabelEncoder. Then, dataset was scaled using StandardScaler.

In order to get the same value every time, the parameter random\_state = 42 was used. The data was split into training and testing sets with a test size of 20% (test\_size=0.2).

### Feature Selection

To perform feature selection SelectKBest with the f\_classif scoring function were used.

Multiple values for k were tested for both linear SVC and standard SVC

The parameter k=4 was selected for the analysis because it achieved one of the best scores while keeping the number of columns low.

The features retained by the selector for k=4 were: month, day\_of\_week, origin\_city\_name and dep\_time

### Initial Model Evaluation (Base SVMs)

All SVM classifiers utilized class\_weight='balanced' to prevent class imbalance, that occurs in the dataset.

#### 4.1. Linear SVC (Baseline)

The initial Linear SVC model showed an accuracy of approximately 0.5857

|  |  |  |
| --- | --- | --- |
| Metric | Class 0.0 (On-time) | Class 1.0 (Delayed) |
| Precision | 0.7118 | 0.4517 |
| Recall | 0.5796 | 0.5961 |
| F1-score | 0.6389 | 0.5140 |

Obraz zawierający diagram, Wykres, linia, tekst

Zawartość wygenerowana przez AI może być niepoprawna.

The model identified about 60% of all truly delayed flights (Recall 1.0 ≈ 0.60). However, when the model predicted a delay, it was correct only 45% of the time (Precision 1.0 = 0.45). The ROC AUC score was approximately 0.623. The low PR AUC (below 0.5) suggested the model struggled to maintain high precision as recall increased, indicating difficulty with class imbalance.

#### SVC RBF Kernel (Baseline)

The SVC using the Radial Basis Function (RBF) kernel achieved an accuracy of 0.58.

|  |  |  |
| --- | --- | --- |
| Metric | Class 0.0 (On-time) | Class 1.0 (Delayed) |
| Precision | 0.7256 | 0.4641 |
| Recall | 0.5774 | 0.6299 |
| F1-SCORE | 0.6443 | 0.5345 |
|  |  |  |

This model was better at catching truly delayed flights (Recall 1.0 = 0.648), which aligns with the goal. It caught 64% of all truly delayed flights. However, like the linear model, when predicting a delay, it was correct only 45% of the time (Precision 1.0 =0.45)

Obraz zawierający diagram, Wykres, linia, tekst

Zawartość wygenerowana przez AI może być niepoprawna.  
  
  
Other Kernels

* **SVC Sigmoid Kernel:** Showed an accuracy of 0.5194. It caught 52% of all truly delayed flights (Recall 1.0 = 0.5224).
* **SVC Polynomial Kernel:** Showed an accuracy of 0.5946. It caught 53% of all truly delayed flights (Recall 1.0 = 0.5344).

### Hyperparameter Optimization using Optuna

Since recall for delayed flights (class 1.0) was the most important metric, hyperparameter tuning was performed using Optuna across three different objectives: maximizing Recall (1.0), maximizing a weighted recall function, and maximizing the F1 score (1.0).

#### Optimized Linear SVC

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Optimization Target | Best Parameters | Accuracy | Recall (Class 1.0) | Recall (Class 0.0) |
| Maximize Recall (1.0) | 'C': 0.9846738873614566, 'penalty': 'l1' | 0.5824 | 0.6131 | 0.5646 |
| Maximize Weighted Recall (0.1\*R0 + 0.9\*R1) | 'C': 0.005934473022625567, 'penalty': 'l1' | 0.5829 | 0.6139 | 0.5649 |
| Maximize F1 Score (1.0) | 'C': 85.98737339212276, 'penalty': 'l1' | 0.5824 | 0.6131 | 0.5646 |

The LinearSVC model optimized for maximized weighted recall score achieved the best recall (approximately 61,4%)  
  
Optimized SVC RBF Kernel

|  |
| --- |
|  |
| Optimization Target | Best Parameters | Accuracy | Recall (Class 1.0) | Recall (Class 0.0) |
| Maximize Recall (1.0) | 'C': 0.0014359718569878096, 'gamma': 0.008113929572637844 | 0.3675 | 1.0 | 0.0 |
| Maximize Weighted Recall (0.4\*R0 + 0.6\*R1) | 'C': 29.026521418263915, 'gamma': 0.0041497957898915935 | 0.555 | 0.7295 | 0.4536 |
| Maximize F1 Score (1.0) | 'C': 92.93816178430156, 'gamma': 0.001444819600787046 | 0.554 | 0.7363 | 0.4481 |

The RBF kernel generally achieved significantly higher recall for the delayed class (Class 1.0) than the LinearSVC variants. However, this high recall came at the cost of poor recall for the on-time class (Class 0.0),.

**The best RBF model based on maximizing the F1 score (1.0) yielded a Recall of approximately 73.6%. This model used parameters C: 92.93816178430156 and gamma: 0.** **001444819600787046.**This model achieved a significantly high Recall for the critical delayed, while still keeping Recall for the not delayed at a decent level.

### Conclusions

The SVM classifiers demonstrated a consistent trade-off where good recall generally meant worse precision.

The goal was to prioritize recall for the delayed flights (Class 1.0).

The RBF model optimized for F1 score gives a balanced view achieving an overall accuracy of 55,4 %.

Ultimately, a model could not be found where both recall values (for delayed and on-time flights) were high. The decision remained to prioritize recall for class 1.0