AIRBUS

CORPORATE PROJECT

Detecting and Predicting Fuel Leaks using Data

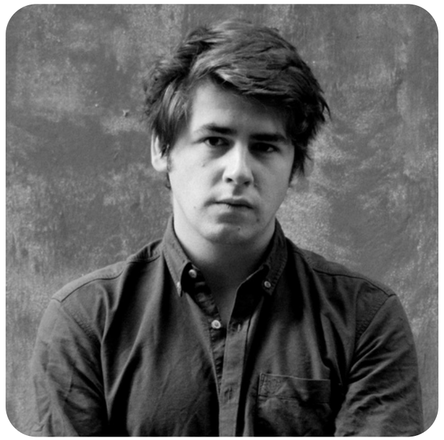
AIRBUS

Our Team

Sarah Awad

Alejandro Danus

Federico Huertas



Ignacio García de Parada

Diego Salazar

ie

AIRBUS

11,164 Plane Crashes

83,772 Deaths

ie

AIRBUS

21% Mechanical Error & Fuel Leaks

17,600 Deaths

ie

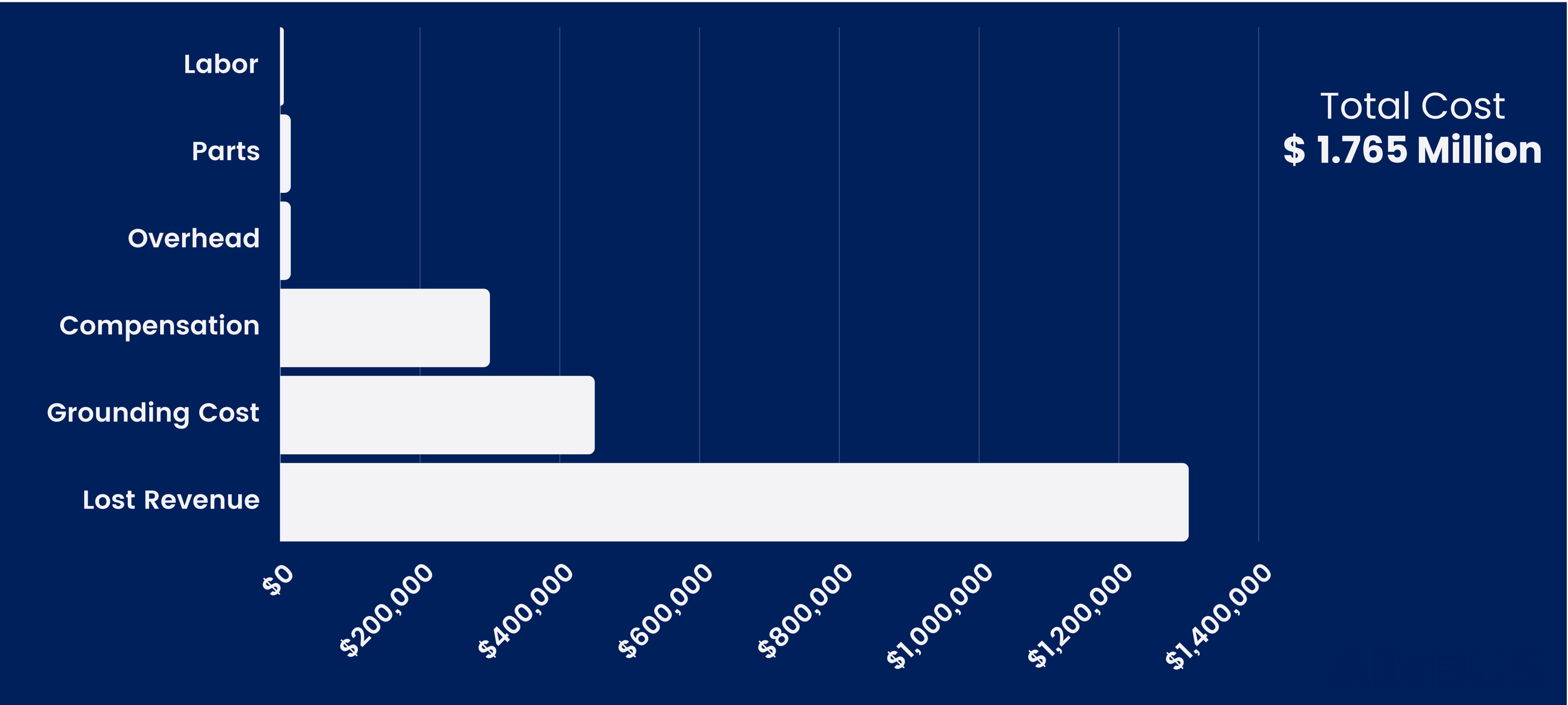
AIRBUS

Boeing

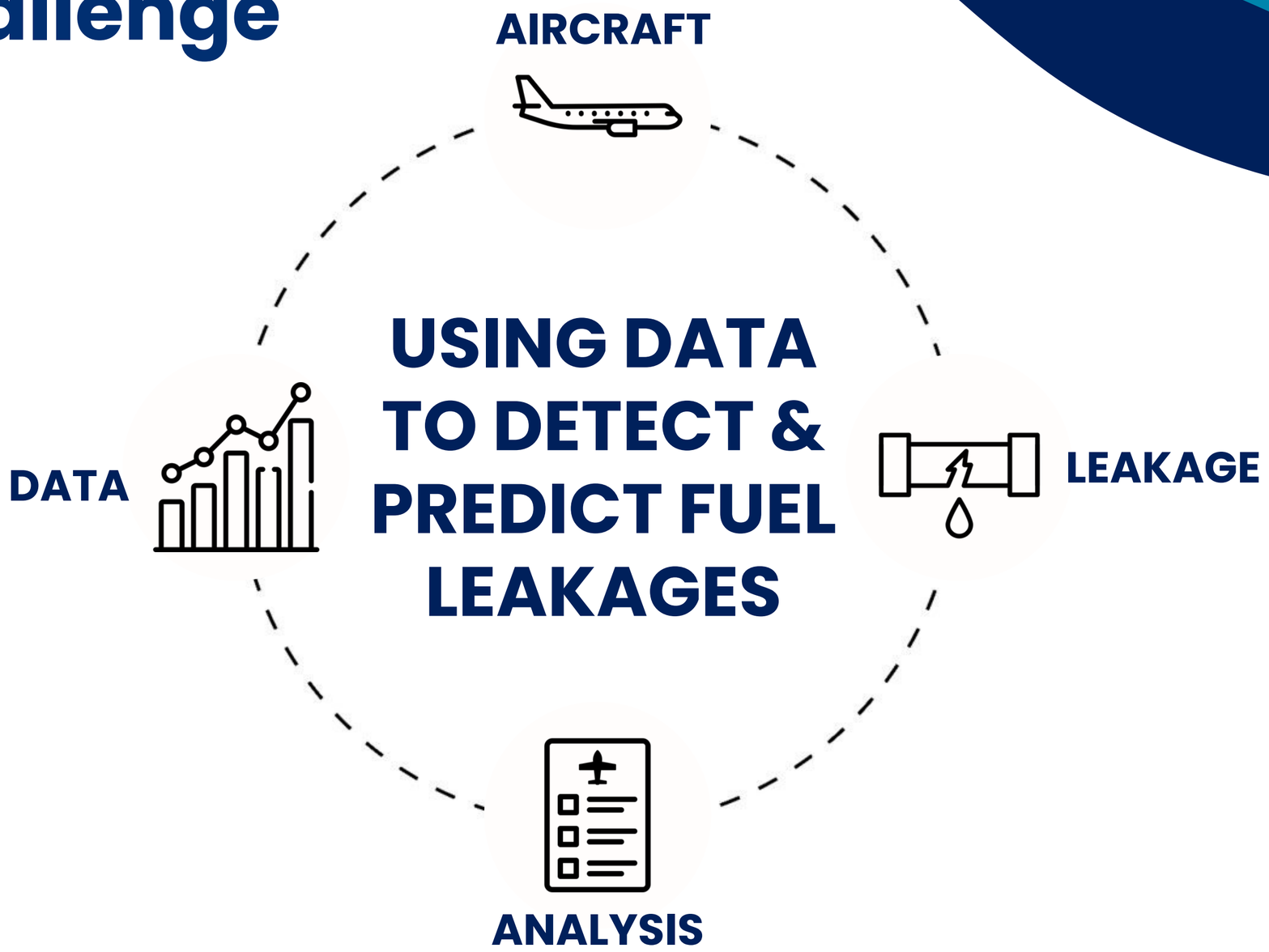
-25% due to 2 crashes

-30 Billion Dollars

Cost Breakdown for 3 days maintenance



Our Challenge



3 MAIN COSTS OF FUEL LEAKS

HUMAN COST

REPUTATIONAL COST

MONETARY COST

Our Solution for Airbus

Logistic Regression Prediction Model

Supervised Model

Proactively identifies fuel leakages

Reduces unnecessary checks

Anomaly Detection Model

Unsupervised Model

Detects anomalies

Reduces the number of false alarms

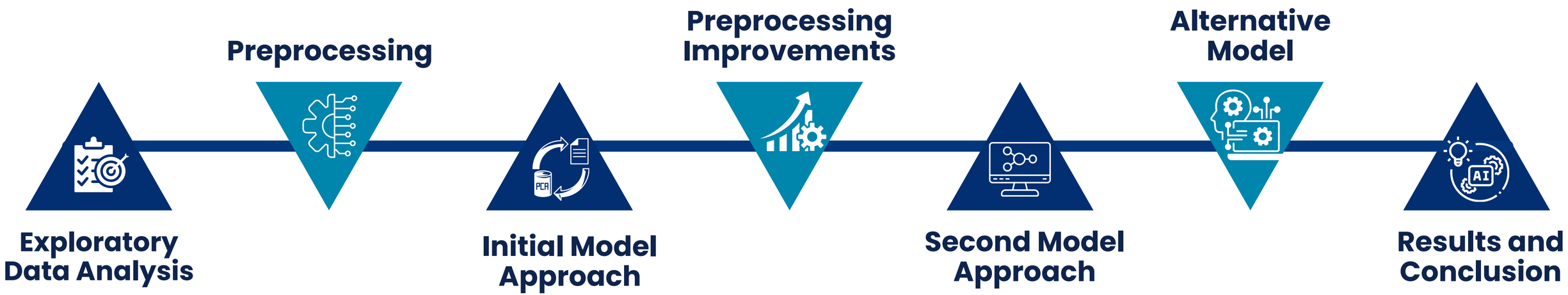
Estimates probability of a leak before it emerges

Locates leaks rapidly through anomalous patterns

ie

AIRBUS

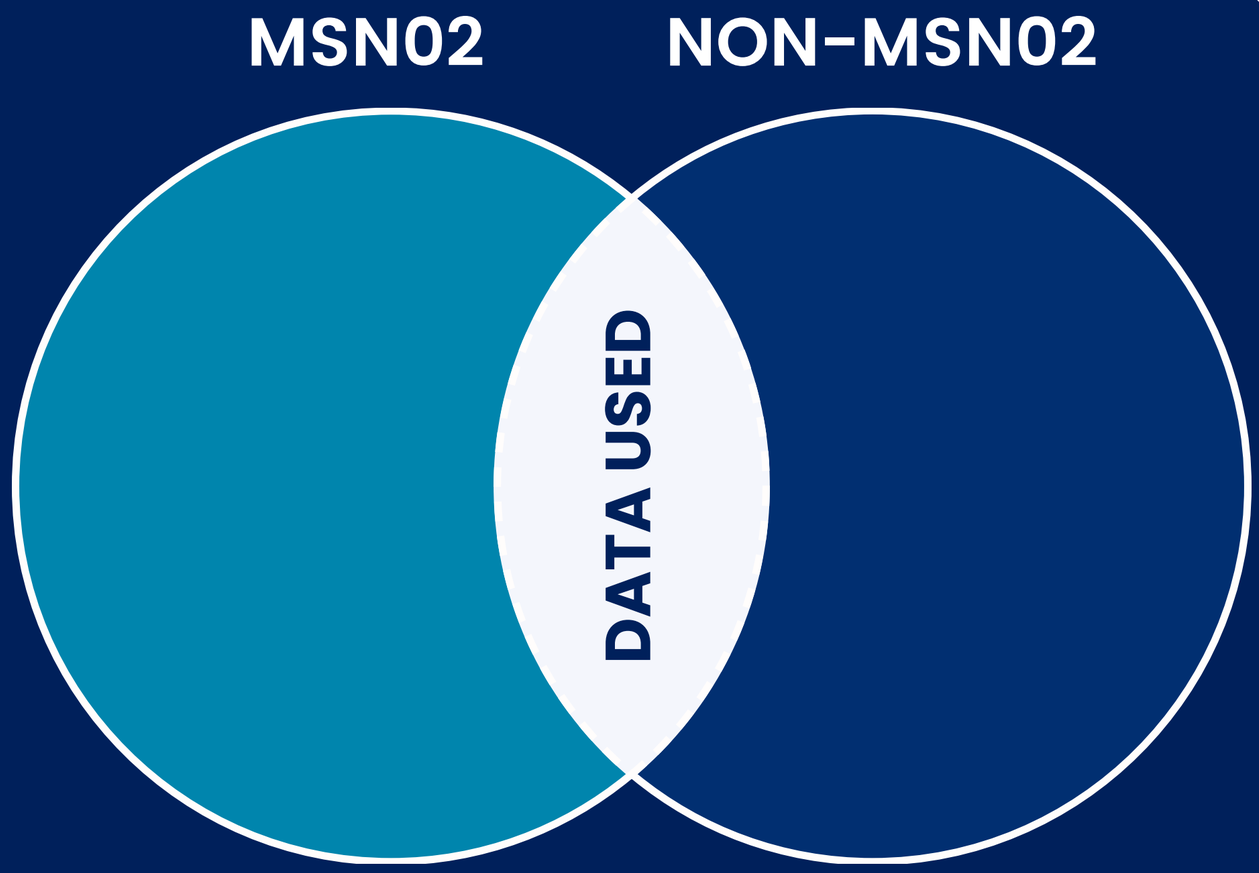
Approach Overview



Exploratory Data Analysis

DATASETS

- 7 commercial - 17 features
- 1 tester - 111 features (MSN02)



Exploratory Data Analysis

WHAT WE DID & FOUND:

Fuel used in each engine had 90% of NULLs (except MSN02)

Phase 8 = Cruise Phase -> More Regular Data

Phase 8 -> Longest Phase

MSN02 only dataset without majority NULLs

Exploratory Data Analysis

HYPOTHESIS

FOB CONTINUOUS DEVIATION vs EXPECTED FOB -> LEAK

For each UTC_TIME period

LEAKAGE FORMULA

LEAKAGE = EXP_FOB - VALUE_FOB

EXP_FOB FORMULA

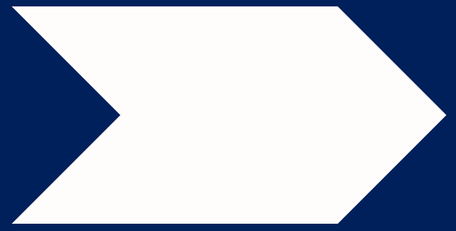
EXP_FOB = VALUE_FOB° - SUM OF TOTAL_FUEL_USED

Preprocessing - Feature Enginiering

From the EDA

MSN 02:

- UTC_TIME
- FLIGHT_PHASE
- FLIGHT
- FW_GEO_ALTITUDE
- VALUE_FOB
- FUEL_USED_1
- FUEL_USED_2
- FUEL_USED_3
- FUEL_USED_4

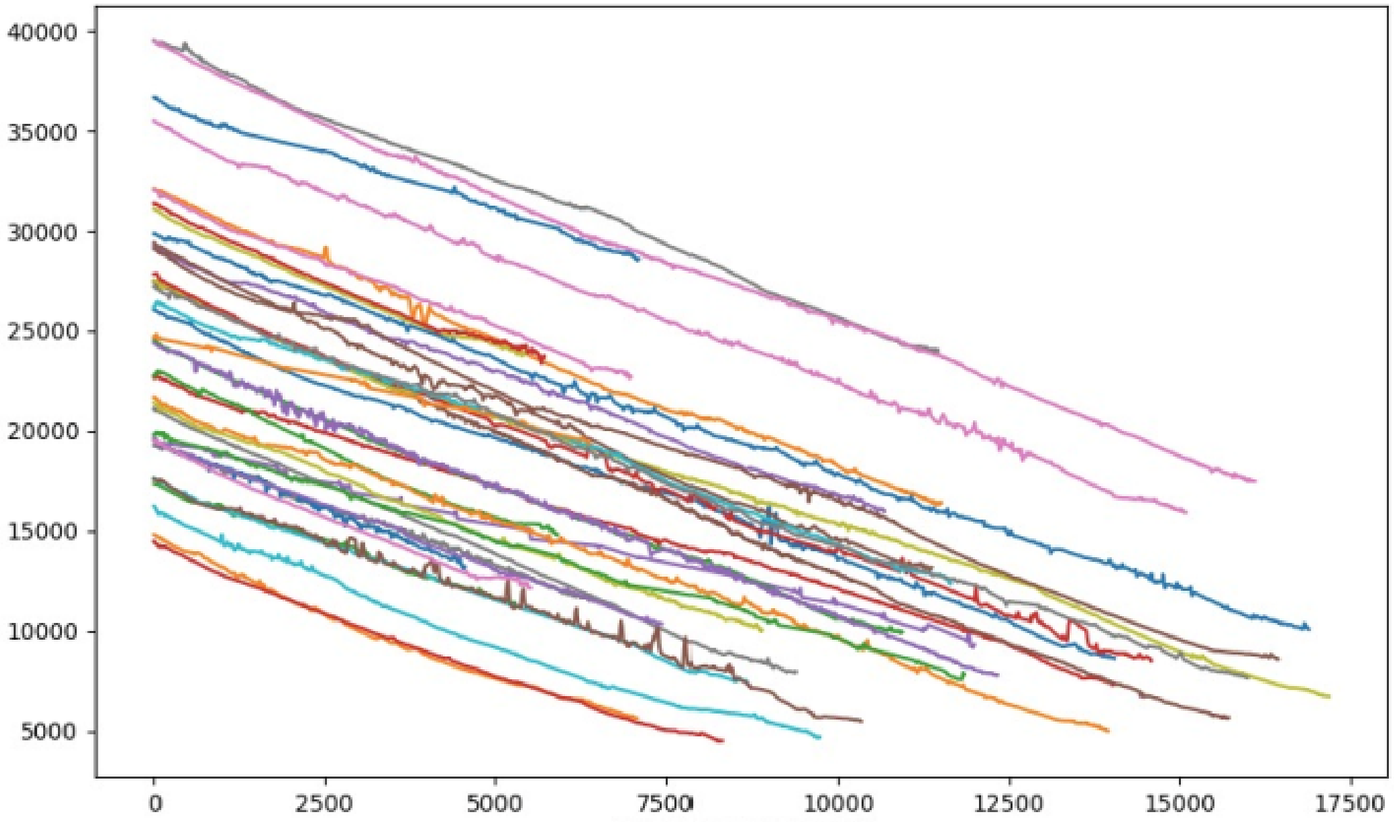


Features created

- SECONDS_PER_FLIGHT
- TOTAL_SECONDS_PER_FLIGHT
- TOTAL_FUEL_USED
- TOTAL_FUEL_USED_DIFF
- VALUE_FOB_DIFF

Preprocessing – Flights overview

Fuel on board vs Seconds per flight



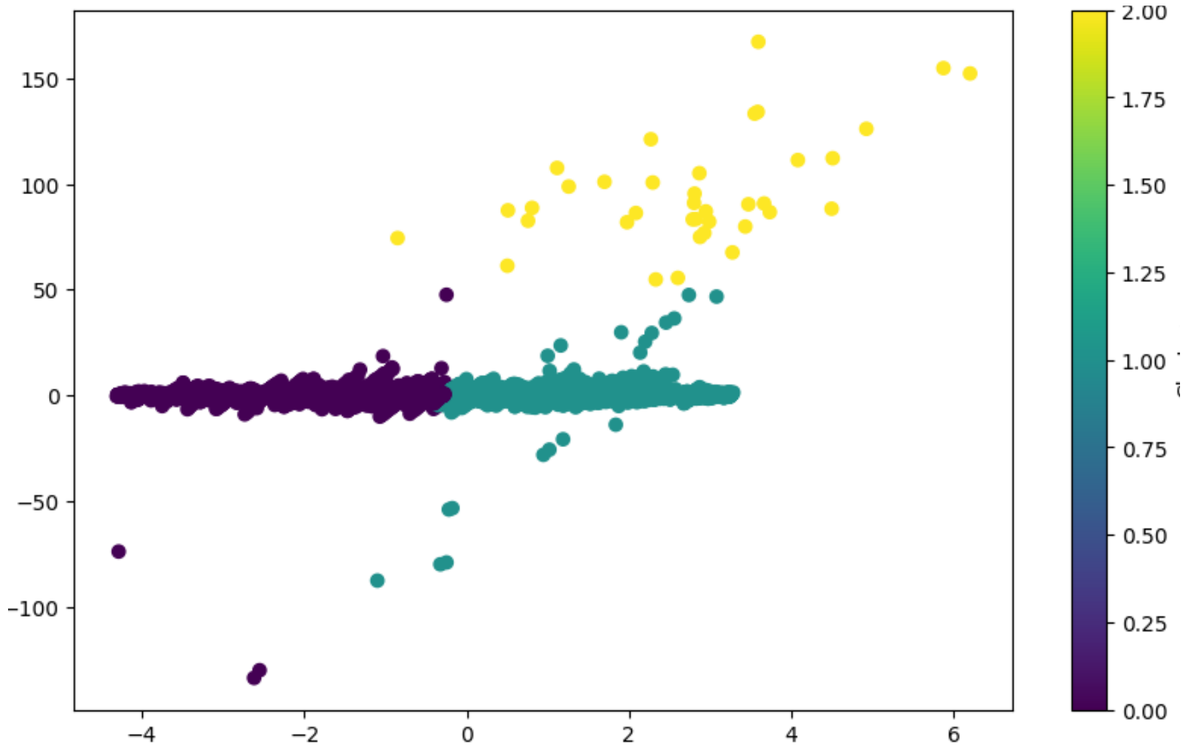
Group the data by flights.

Filter: flights > 30 minutes

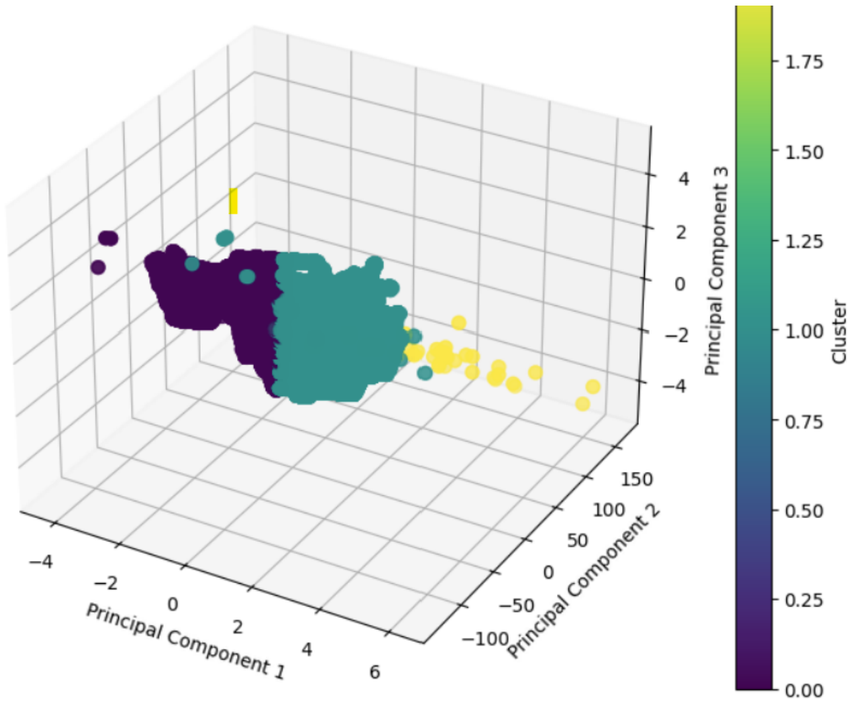
Cluster the Data?

Preprocessing - PCA with K-means

2 components



3 components



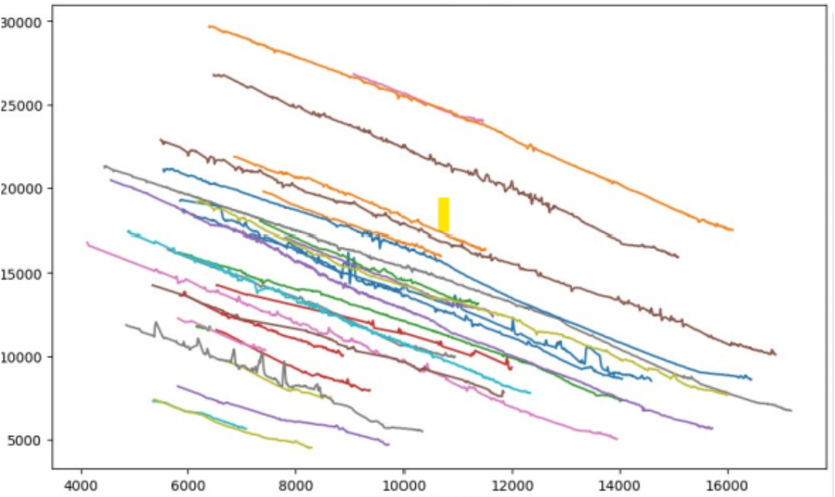
2 components and 3 components.
3 clusters.

Preprocessing - PCA with K-means

FOB vs Seconds per Flight

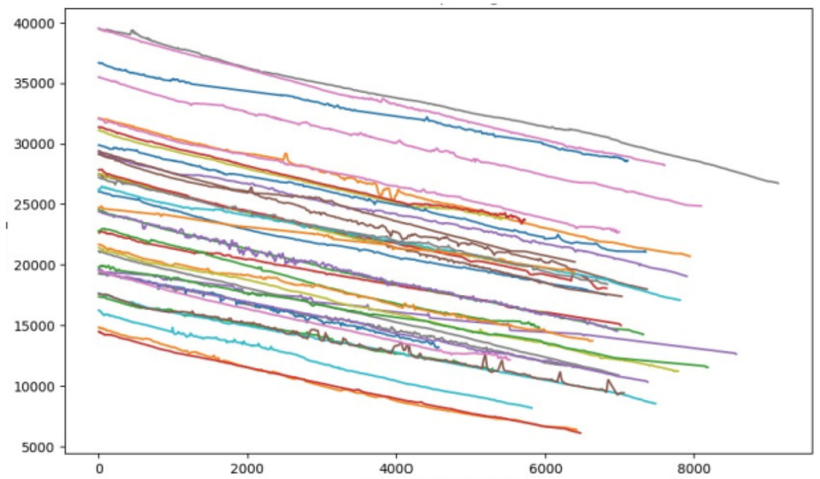
Cluster 0

Fuel on board vs Seconds per flight



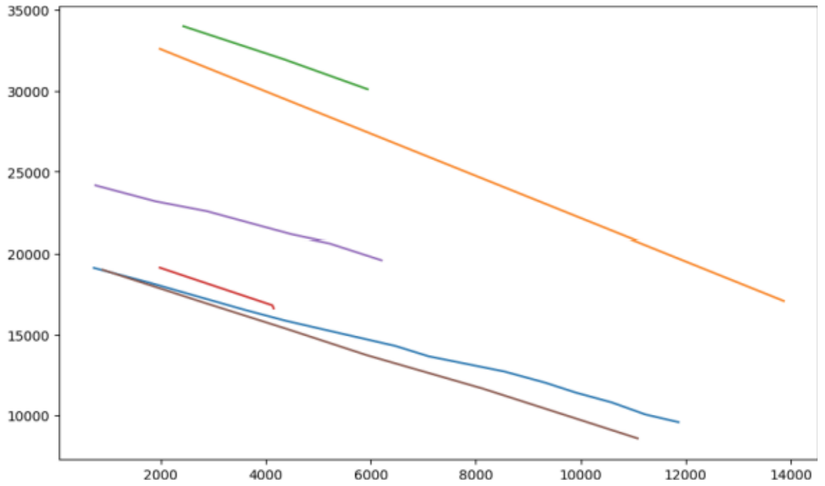
Cluster 1

Fuel on board vs Seconds per flight



Cluster 2

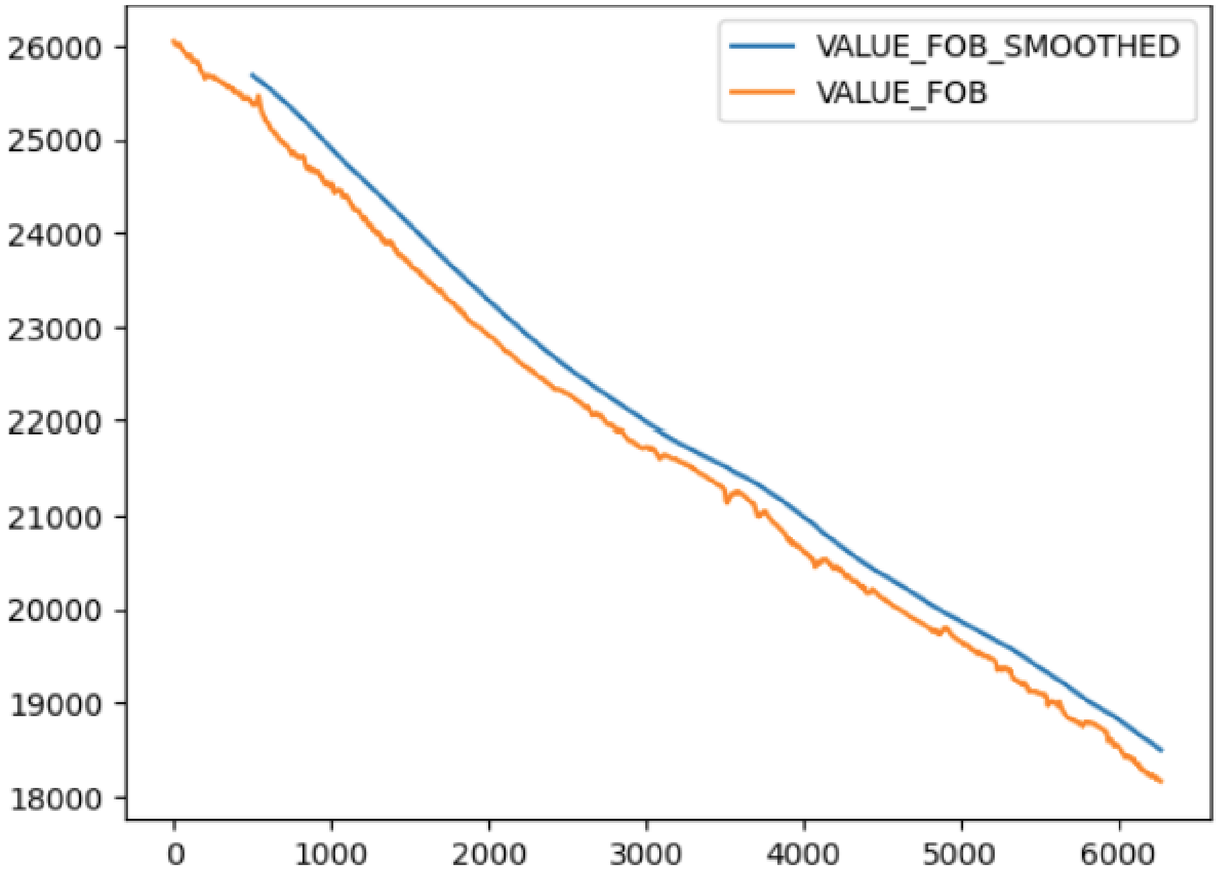
Fuel on board vs Seconds per flight



Cluster 1 flights are more similar.

Preprocessing - Smoothing

Fuel on board vs Smoothed Fuel on board



Outliers were dropped

Smoothing the data

Creating variables:

VALUE_FOB_SMOOTHED &

VALUE_FOB_SMOOTHED_DIFF

Preprocessing - Simulated Leakage

Goal:

Make a classification model

Steps:

Created 2 copies of our cleaned MSN02.

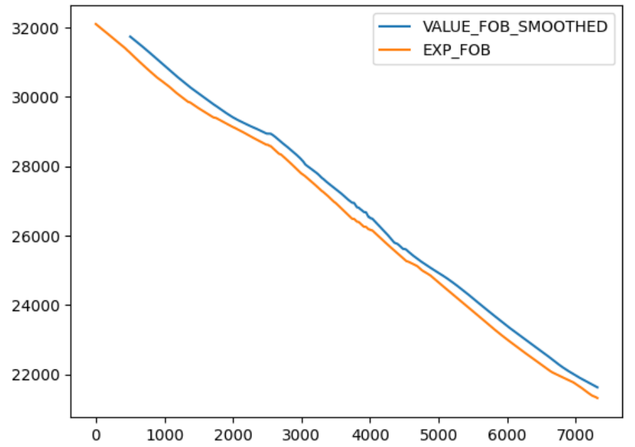
1 maintained as it was. --> Leakage = 0

3 groups of Random Flights.

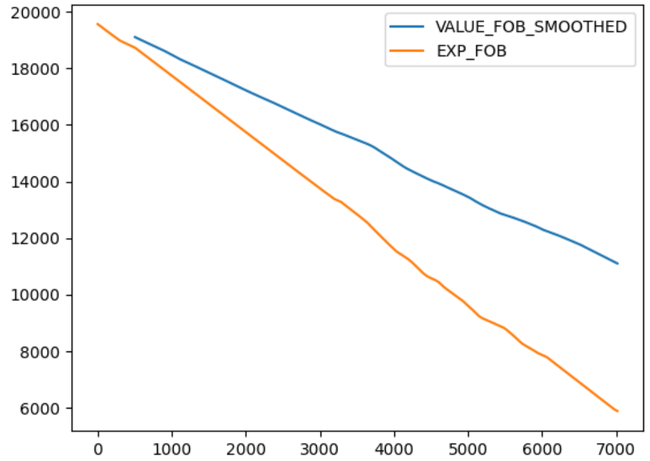
Simulated leakage 0.25, 0.75, 0.9.

1 we simulated leakeges. --> Leakage = 1

Without Leakage



With Leakage



First Model Approach

Logistic Regression Prediction Model

Logistic Regression **without** Fuel Used

RECALL = 53%

Logistic Regression **with** Fuel Used

RECALL = 90%

MSN_02 Key Feature Analysis

MSN02 CM

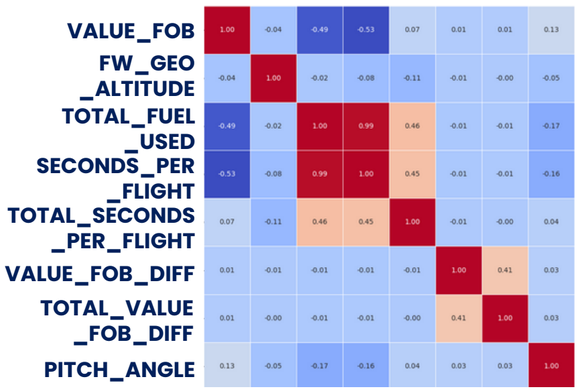
SELECTED FEASIBLE FEATURES ANALYSIS

	ENGINE_RUNNING_1	ENGINE_RUNNING_2	ENGINE_RUNNING_3	ENGINE_RUNNING_4	FUEL_FLOW_1	FUEL_FLOW_2	FUEL_FLOW_3	FUEL_FLOW_4	FUEL_PITCH	FUEL_ROLL	FUEL_USED_1	FUEL_USED_2	FUEL_USED_3	FUEL_USED_4	FW_GEO_ALTITUDE	PITCH_ANGLE	ROLL_ANGLE	VALUE_FOB
ENGINE_RUNNING_1	1.00	0.87	0.83	0.67	0.44	0.42	0.42	0.40	0.14	-0.08	0.08	0.07	0.05	0.06	0.21	0.22	-0.01	-0.06
ENGINE_RUNNING_2	0.87	1.00	0.95	0.65	0.43	0.45	0.45	0.40	0.14	-0.08	0.09	0.07	0.04	0.05	0.22	0.23	-0.01	-0.06
ENGINE_RUNNING_3	0.83	0.95	1.00	0.68	0.43	0.45	0.47	0.42	0.15	-0.09	0.11	0.10	0.07	0.06	0.23	0.24	-0.01	-0.07
ENGINE_RUNNING_4	0.67	0.65	0.68	1.00	0.26	0.25	0.27	0.58	0.22	-0.07	0.03	0.03	0.01	0.22	0.21	0.08	-0.01	-0.09
FUEL_FLOW_1	0.44	0.43	0.43	0.26	1.00	0.92	0.93	0.81	0.43	-0.20	0.06	0.05	0.04	-0.01	0.35	0.67	-0.02	0.05
FUEL_FLOW_2	0.42	0.45	0.45	0.25	0.92	1.00	0.96	0.77	0.42	-0.19	0.06	0.06	0.04	-0.01	0.36	0.67	-0.01	0.07
FUEL_FLOW_3	0.42	0.45	0.47	0.27	0.93	0.96	1.00	0.78	0.43	-0.16	0.06	0.06	0.04	-0.02	0.38	0.68	-0.01	0.05
FUEL_FLOW_4	0.40	0.40	0.42	0.58	0.81	0.77	0.78	1.00	0.49	-0.15	0.01	0.01	-0.01	0.13	0.38	0.51	-0.01	0.02
FUEL_PITCH	0.14	0.14	0.15	0.22	0.43	0.42	0.43	0.49	1.00	-0.15	-0.08	-0.08	-0.08	0.01	0.21	0.52	-0.01	0.01
FUEL_ROLL	-0.08	-0.08	-0.09	-0.07	-0.20	-0.19	-0.16	-0.15	-0.15	1.00	-0.11	-0.11	-0.10	-0.10	-0.15	-0.19	0.09	0.03
FUEL_USED_1	0.08	0.09	0.11	0.03	0.06	0.06	0.06	0.01	-0.08	-0.11	1.00	0.99	0.97	0.82	0.14	0.05	-0.00	-0.52
FUEL_USED_2	0.07	0.07	0.10	0.03	0.05	0.06	0.06	0.01	-0.08	-0.11	0.99	1.00	0.98	0.82	0.15	0.05	-0.00	-0.51
FUEL_USED_3	0.05	0.04	0.07	0.01	0.04	0.04	0.04	-0.01	-0.08	-0.10	0.97	0.98	1.00	0.82	0.14	0.04	-0.00	-0.52
FUEL_USED_4	0.06	0.05	0.06	0.22	-0.01	-0.01	-0.02	0.13	0.01	-0.10	0.82	0.82	0.82	1.00	0.15	-0.02	-0.00	-0.49
FW_GEO_ALTITUDE	0.21	0.22	0.23	0.21	0.35	0.36	0.38	0.38	0.21	-0.15	0.14	0.15	0.14	0.15	1.00	0.34	-0.02	-0.12
PITCH_ANGLE	0.22	0.23	0.24	0.08	0.67	0.67	0.68	0.51	0.52	-0.19	0.05	0.05	0.04	-0.02	0.34	1.00	0.00	-0.00
ROLL_ANGLE	-0.01	-0.01	-0.01	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	0.09	-0.00	-0.00	-0.00	-0.00	-0.02	0.00	1.00	0.02
VALUE_FOB	-0.06	-0.06	-0.07	-0.09	0.05	0.07	0.05	0.02	0.01	0.03	-0.52	-0.51	-0.52	-0.49	-0.12	-0.00	0.02	1.00

Filtered MSN02 CM

FUEL_FLOW_AVE ANALYSIS

PITCH_ANGLE ANALYSIS



FUEL_FLOW_AVE

PITCH_ANGLE

Implementation for Baseline Model Improvements

Key Features Random Forest Model Scores

FUEL_FLOW_AVE

MAPE = 1.41%

PITCH_ANGLE

MAPE ≃ 0

Implementation in Logistic Regression Model Scores

With FUEL_FLOW_AVE

RECALL= 75.21%

With PITCH_ANGLE

RECALL= 59.12%

Second Model Approach

Logistic Regression Prediction Model

Logistic Regression **without** Fuel Used & **with** Fuel Flow Average

RECALL = 75%

Additional Model Approach: LSTM Autoencoder

Why LSTM?

- Unsupervised learning methods
- Identifies data patterns and flags anomalies

Feature addition and model building

1. Percentage variation of Fuel on Board and Total Fuel Used
2. Polynomial representation of Fuel on Board
3. Difference between real and expected Fuel on Board

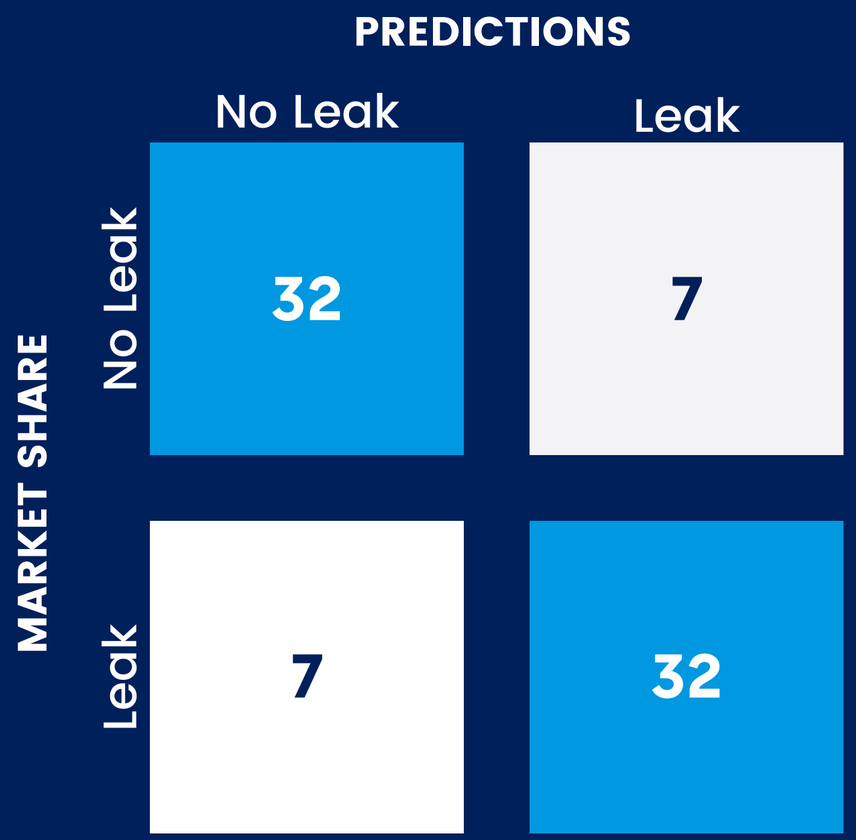
Data preparation and model training

Scaling numeric features

Structuring data in sequences (30 mins)

LSTM Results

Results using msn_37 dataset (including artificial leaks)



Considerations:

1. Incorporate new features
2. Increase training set size with more leak examples
3. Experimenting with different configurations
4. Identify patterns to define optimal classification thresholds

82.5% Recall

All Model Results

	MODELS		
Metrics	**Logistic Regression without Fuel Used**	**Logistic Regression with Fuel Used**	**LSTM AUTOENCODER**
Recall	**75.21%**	**90.02%**	**82%**
% of Leaks Predicted	5.56%	3.63%	N/A

Learnings & Next Steps

Improvements

Improve feature engineering and preprocessing

Add features correlated with fuel used, not included in the other datasets

Garbage in garbage out!

Additional Phases

Extend our analysis to additional phases

Sensor data will have to be more reliable

Use weight as a measurement of fuel amount

Complex Models

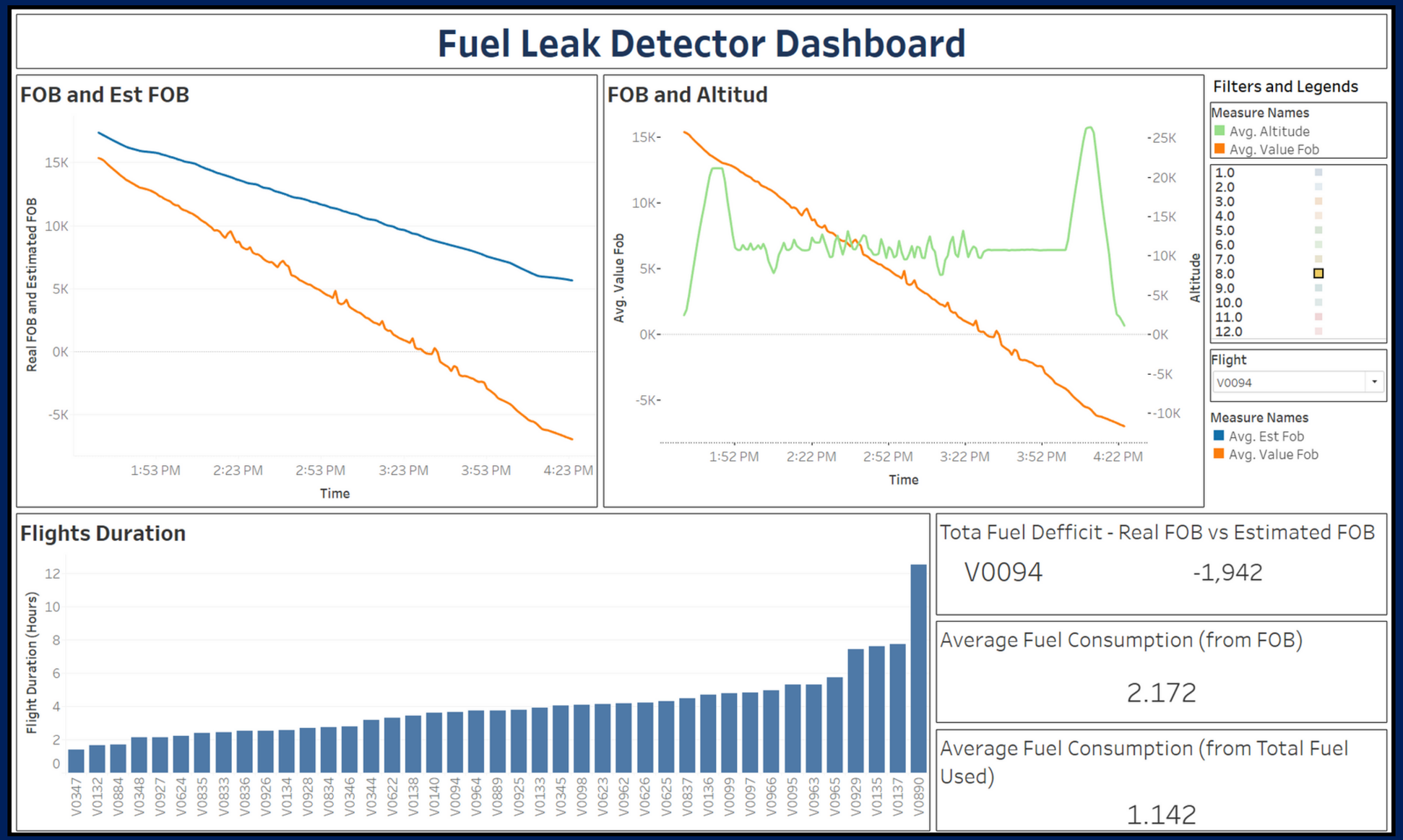
Recurrent Neural Networks

RNNs can capture long-sequence patterns

Local Outlier Factor

Finds data that deviates from the norm

Dashboard Link



THANK YOU

All Model Results

	MODELS		
Metrics	**Logistic Regression without Fuel Used**	**Logistic Regression with Fuel Used**	**LSTM AUTOENCODER**
Recall	**75.21%**	**90.02%**	**82%**
Precision			**82%**
Accuracy			**82%**
F1 - Score			**82%**

ie

AIRBUS

1st Logistic Regression Code Results

Including FUEL_USED

```
Logistic Regression Accuracy: 0.9022
Logistic Regression F1 Score: 0.9005
Logistic Regression Recall: 0.8872
Logistic Regression Precision: 0.9143
```

Not including FUEL_USED

```
Best Parameters: {'C': 1, 'penalty': 'l2'}
Logistic Regression Accuracy: 0.4980
Logistic Regression F1 Score: 0.5148
Logistic Regression Recall: 0.5289
Logistic Regression Precision: 0.5015
```

2nd Logistic Regression Code Results

Including FUEL_USED

```
Best Parameters: {'C': 0.001, 'penalty': 'none'}
Logistic Regression Accuracy: 0.9314
Logistic Regression F1 Score: 0.9290
Logistic Regression Recall: 0.9019
Logistic Regression Precision: 0.9579
```

Not including FUEL_USED

```
Best Parameters: {'C': 0.001, 'penalty': 'l2'}
Logistic Regression Accuracy: 0.4942
Logistic Regression F1 Score: 0.5952
Logistic Regression Recall: 0.7521
Logistic Regression Precision: 0.4925
```

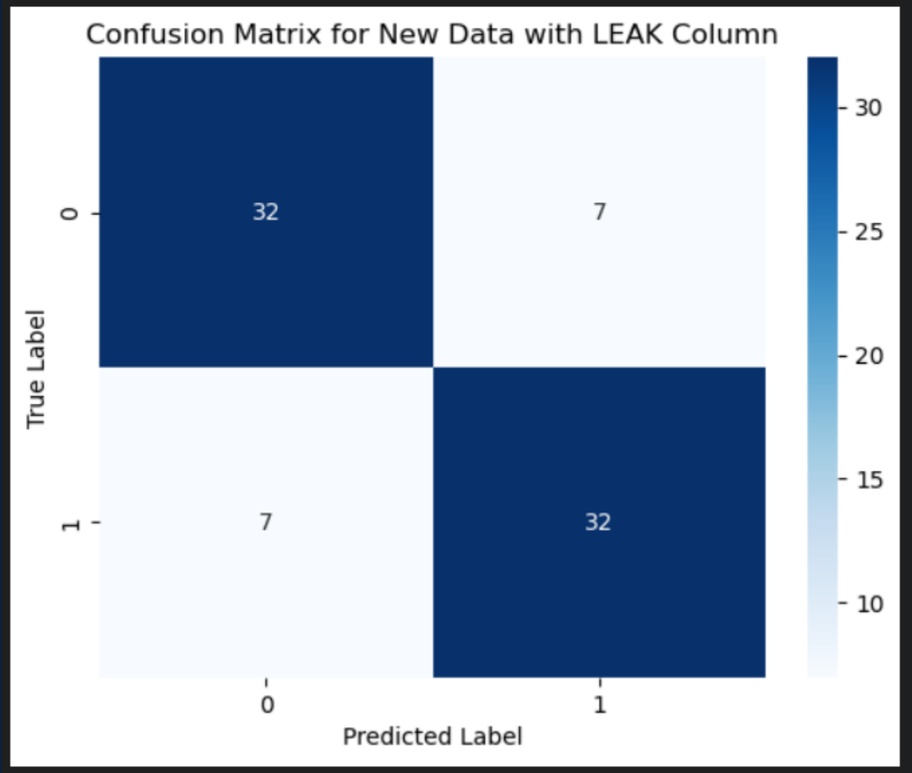
```python
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, f1_score, recall_score, precision_score
param_grid = {
    'C': [0.001, 0.01, 0.1, 1, 10, 100],
    'penalty': ['l2', 'none']
}
log_reg = LogisticRegression(solver='lbfgs', random_state=42)
grid_search = GridSearchCV(log_reg, param_grid, cv=5, scoring='f1', n_jobs=-1)
grid_search.fit(X_train_normalized, y_train)
best_params = grid_search.best_params_
best_log_reg = grid_search.best_estimator_
y_val_pred_log = best_log_reg.predict(X_val_normalized)
accuracy_log = accuracy_score(y_val, y_val_pred_log)
f1_log = f1_score(y_val, y_val_pred_log)
recall_log = recall_score(y_val, y_val_pred_log)
precision_log = precision_score(y_val, y_val_pred_log)
print("Best Parameters:", best_params)
print(f"Logistic Regression Accuracy: {accuracy_log:.4f}")
print(f"Logistic Regression F1 Score: {f1_log:.4f}")
print(f"Logistic Regression Recall: {recall_log:.4f}")
print(f"Logistic Regression Precision: {precision_log:.4f}")
```

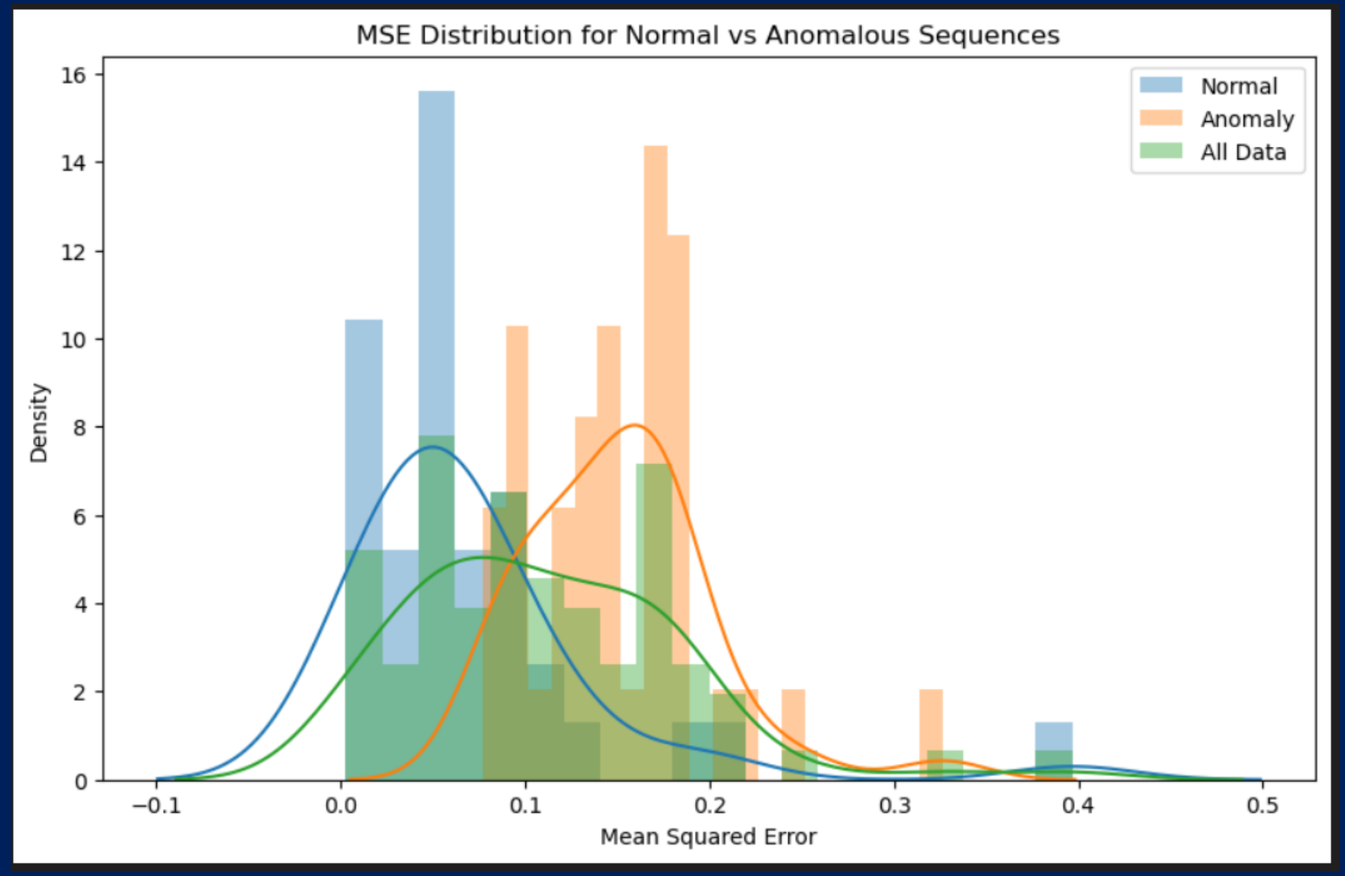
XGBoost Code Results

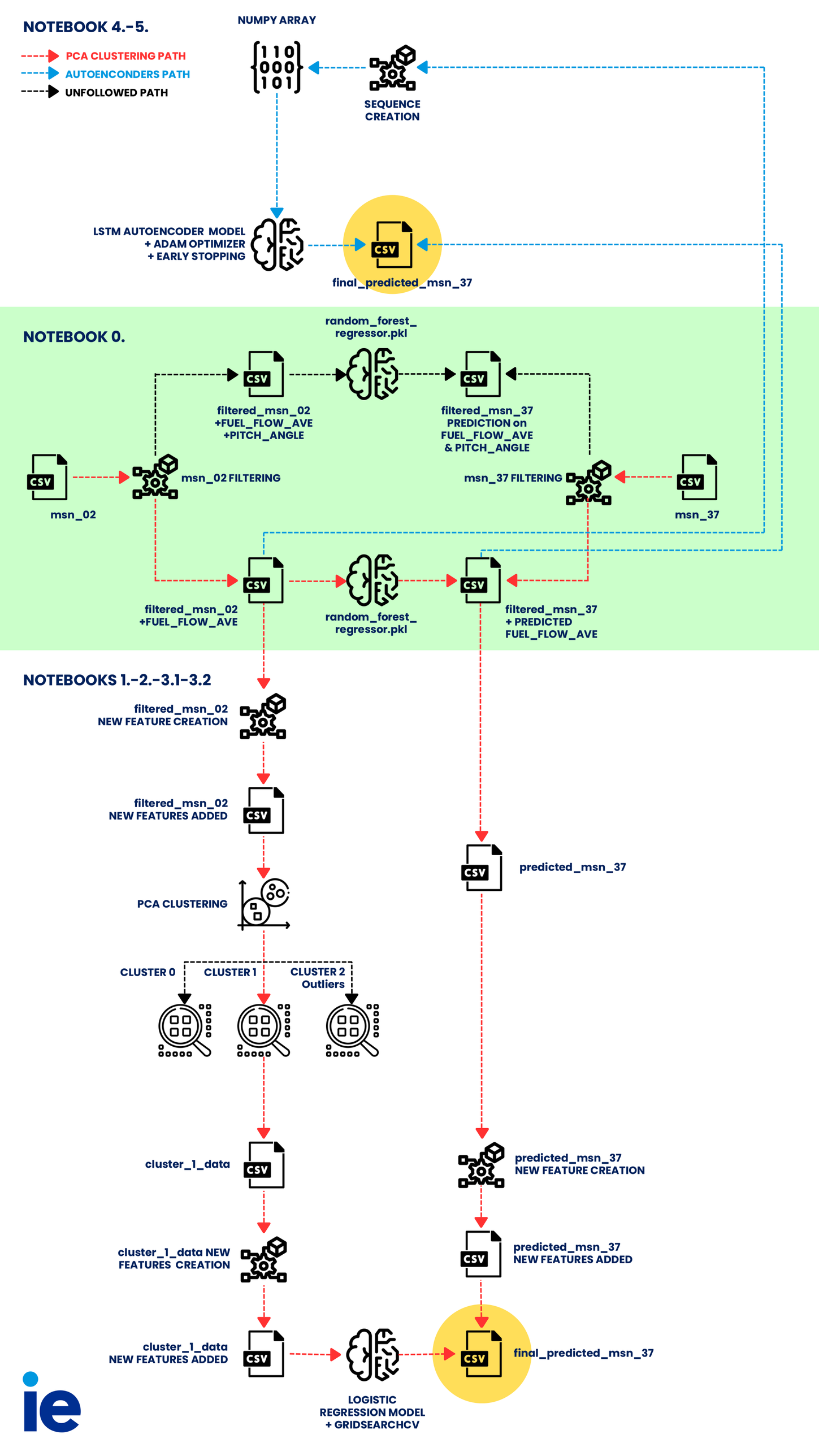
```
XGBoost Accuracy: 0.3753995157384988
XGBoost F1 Score: 0.3761849487328303
XGBoost Recall: 0.3739782671410712 4
XGBoost  Precision: 0.3784178262138756 5
```

LSTM Code Results

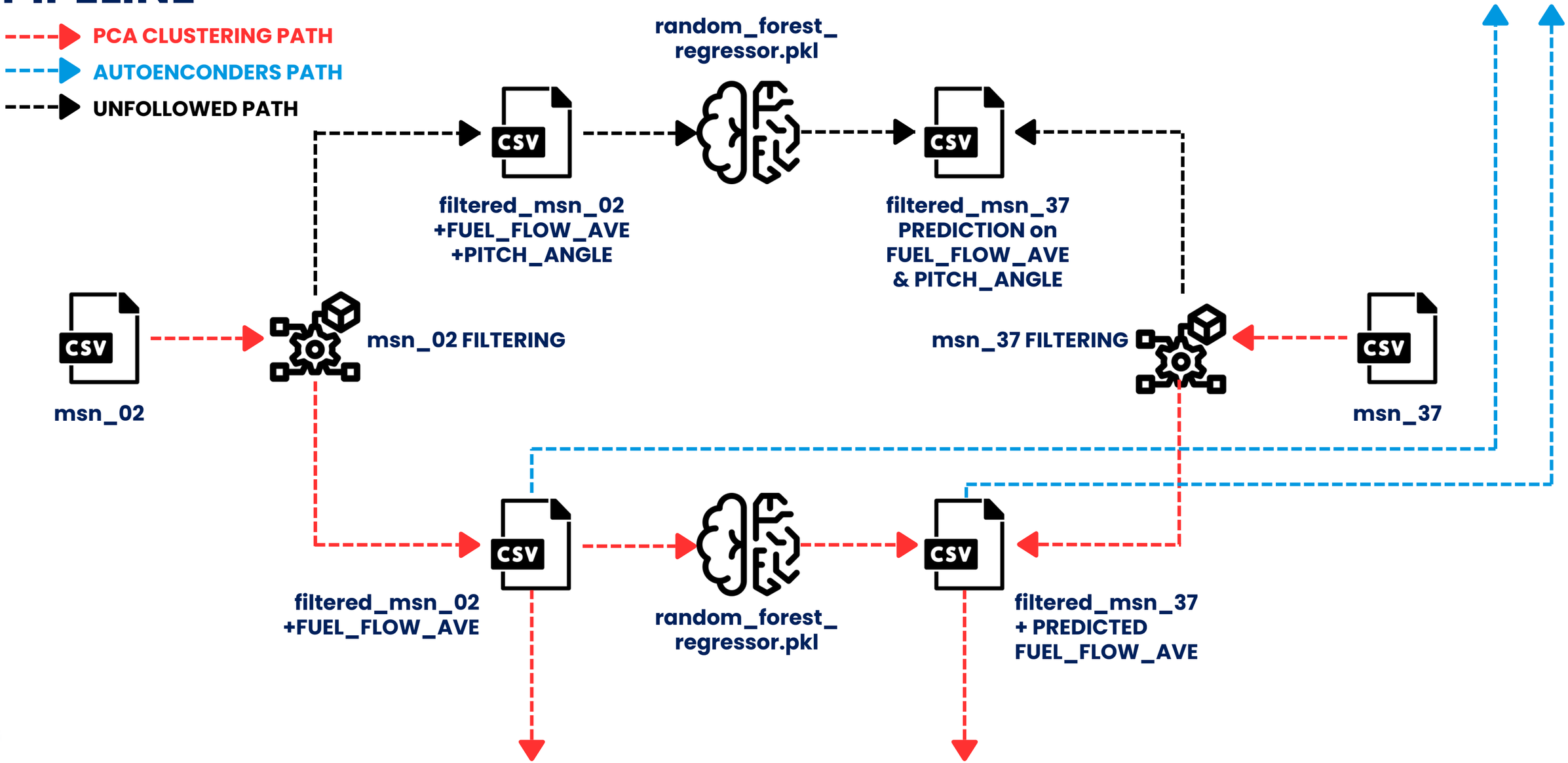
	precision	recall	f1-score	support
0	0.82	0.82	0.82	39
1	0.82	0.82	0.82	39
accuracy			0.82	78
macro avg	0.82	0.82	0.82	78
weighted avg	0.82	0.82	0.82	78



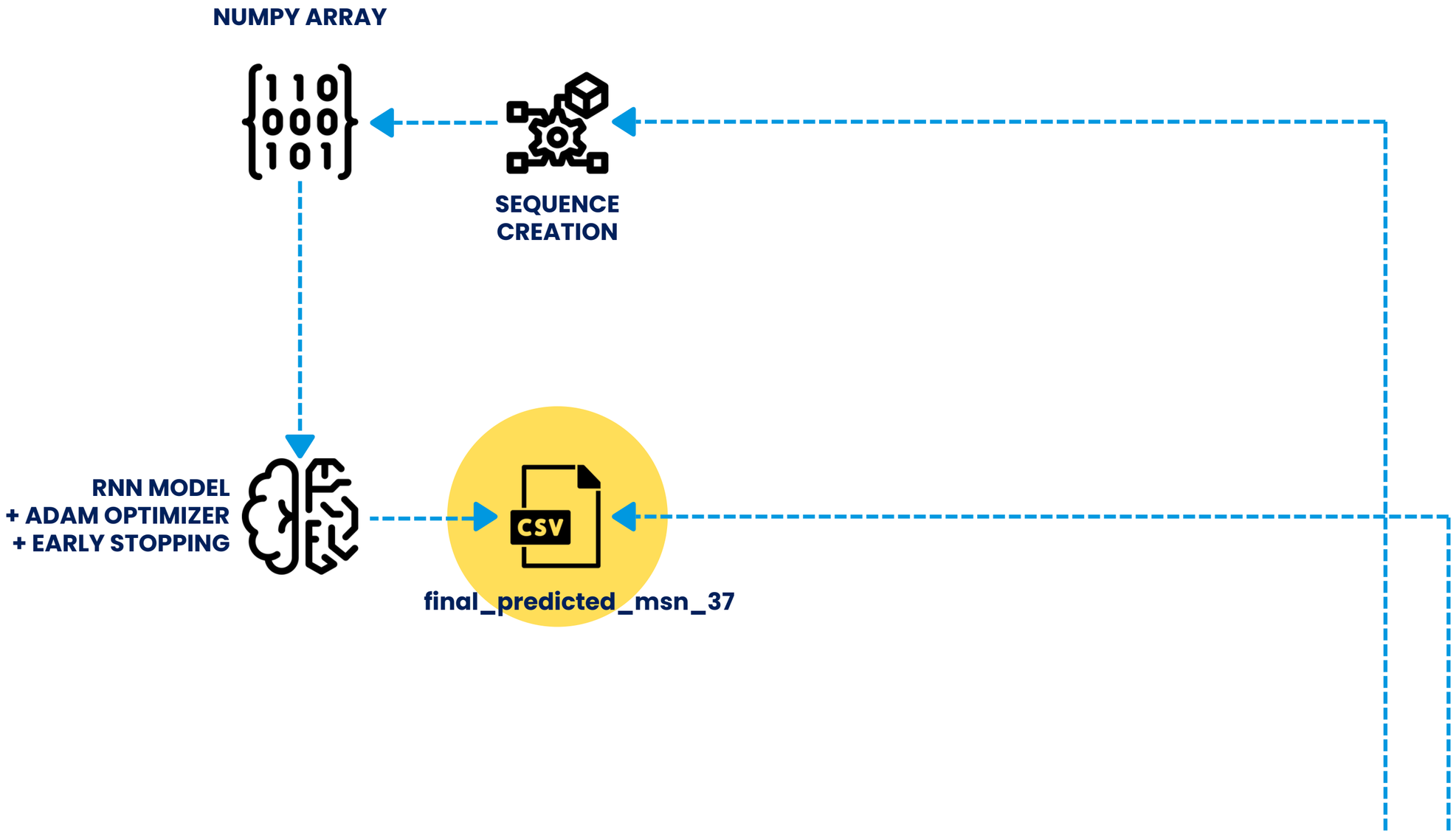



NOTEBOOK 4.-5.
PCA CLUSTERING PATH
AUTOENCONDERS PATH
UNFOLLOWED PATH
NUMPY ARRAY
SEQUENCE CREATION
LSTM AUTOENCODER MODEL
+ ADAM OPTIMIZER
+ EARLY STOPPING
CSV
final_predicted_msn_37
NOTEBOOK 0.
random_forest_ regressor.pkl
filtered_msn_02 +FUEL_FLOW_AVE +PITCH_ANGLE
filtered_msn_37 PREDICTION on FUEL_FLOW_AVE & PITCH_ANGLE
msn_02
msn_02 FILTERING
msn_37 FILTERING
msn_37
filtered_msn_02 +FUEL_FLOW_AVE
random_forest_ regressor.pkl
filtered_msn_37 + PREDICTED FUEL_FLOW_AVE
NOTEBOOKS 1.-2.-3.1-3.2
filtered_msn_02 NEW FEATURE CREATION
filtered_msn_02 NEW FEATURES ADDED
predicted_msn_37
PCA CLUSTERING
CLUSTER 0
CLUSTER 1
CLUSTER 2 Outliers
cluster_1_data
predicted_msn_37 NEW FEATURE CREATION
cluster_1_data NEW FEATURES CREATION
predicted_msn_37 NEW FEATURES ADDED
cluster_1_data NEW FEATURES ADDED
final_predicted_msn_37
LOGISTIC REGRESSION MODEL + GRIDSEARCHCV

FEATURE AUGMENTATION PIPELINE



AUTOENCODERS PATH



PCA CLUSTERING PATH

filtered_msn_02
NEW FEATURE CREATION

filtered_msn_02
NEW FEATURES ADDED

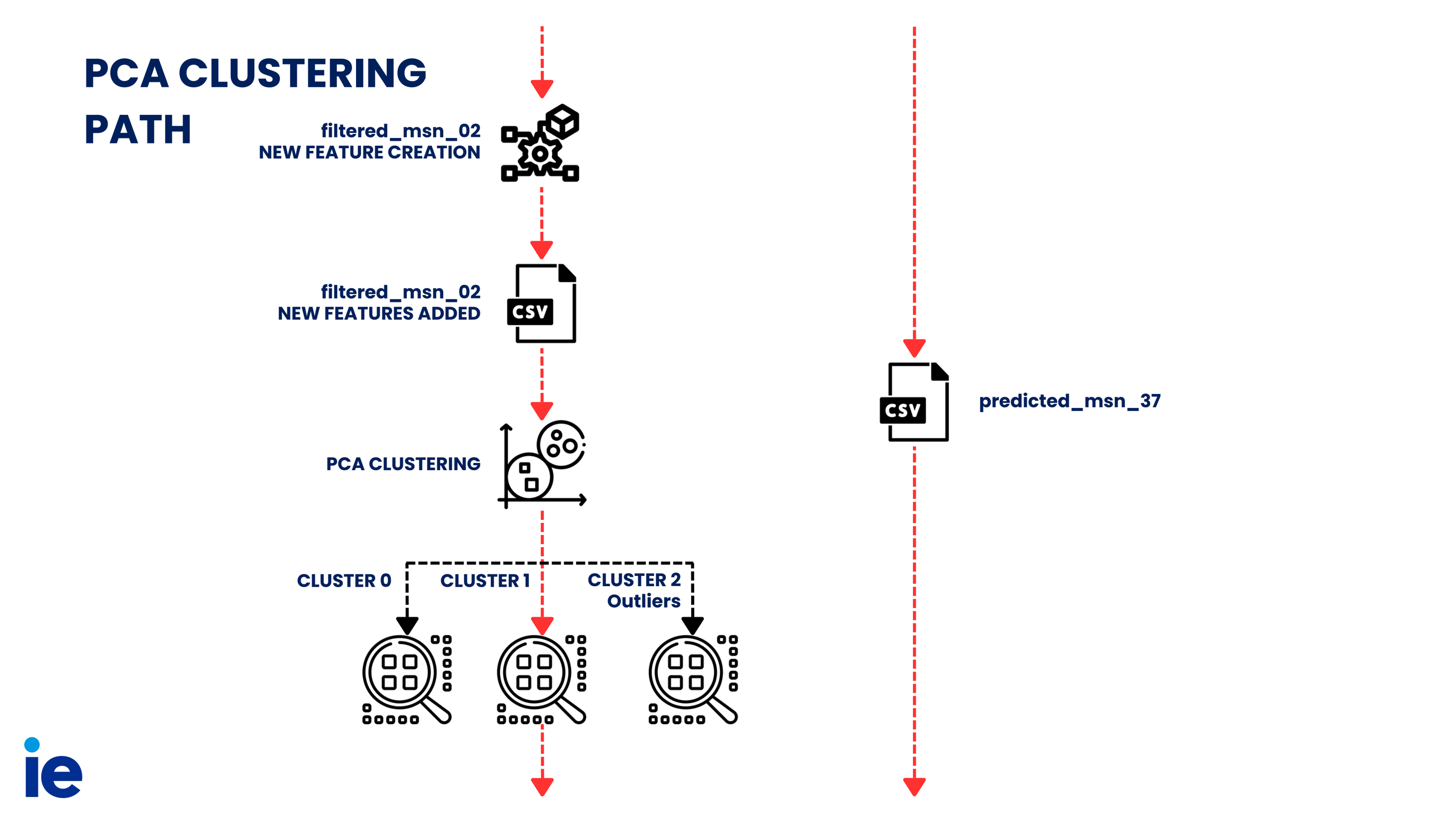
PCA CLUSTERING

CLUSTER 0

CLUSTER 1

CLUSTER 2
Outliers

predicted_msn_37



PCA CLUSTERING PATH