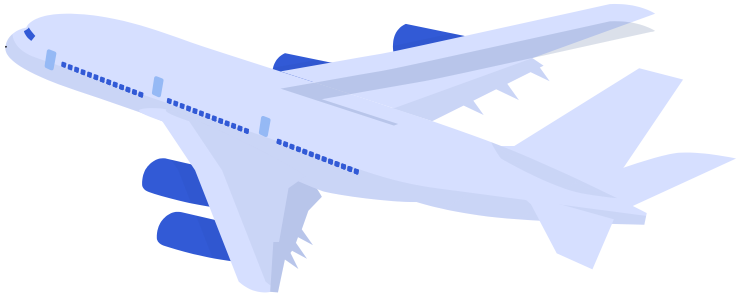
From Delays to Decisions: Predictive Insights for U.S. Airlines

Valeria Breton, Lilie Catania, and Liza Scott



Presentation Agenda

1. ***Executive Summary***
   Business question, dataset introduction, and project objectives

2. ***Exploratory Analysis***
   Using Data Exploration to Shape Business Questions

3. ***Regression Modeling***
   Includes bivariate and multivariate models

4. ***Classification Modeling***
   Binary and multiclass model comparisons

5. ***Summarized Results***
   Recap of model performances, lessons learned, and implications

6. ***Conclusion & Next Steps***
   Final takeaways, project limitations, and recommendations for future work

1

Executive

A high-level data overview

How can flight delay predictions help airlines optimize operations and improve customer experience?

Think— staffing investments, flight scheduling adjustments, customer communication strategies, and overall workflow efficiency.

Note: The dataset was sourced from Kaggle. No data collection was conducted by our team.

Dataset Introduction

2015 Flight Delays and Cancellations
Collected and published by the U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics

Raw Data

Fields: 31

Records: 5.28 Million

Clean Data

Fields: 27

Records: 2,443,491

Training (70%)
Fields: 27
Records: 1,710,443

Validation (15%)
Fields: 27
Records: 366,524

Testing (15%)
Fields: 27
Records: 366,524

Executive Summary: Key Variables

Departure Delay

The difference between the scheduled and actual departure time of a flight, measured in min.

Day of the Week

Indicates which day (Monday-Sunday) the flight is scheduled to depart.

Airline

The carrier operating the flight, identified by airline code.

Distance

The total distance of the flight route, measured in miles.

Origin/Destination Airport

The departure and arrival airports for the flight, identified by their respective IATA codes.

Weather Delay

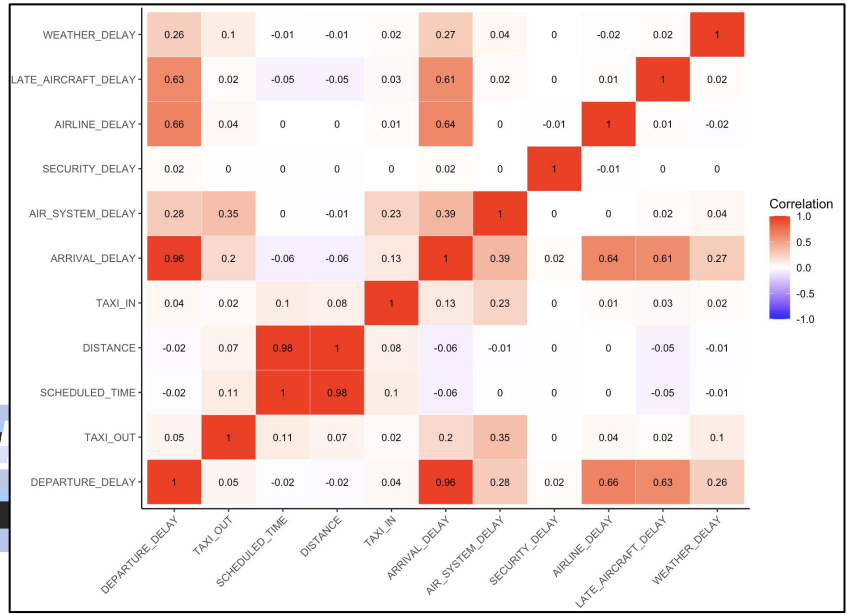
The amount of delay (in minutes) attributed to weather-related issues affecting flight operations.

2

Exploratory Analysis

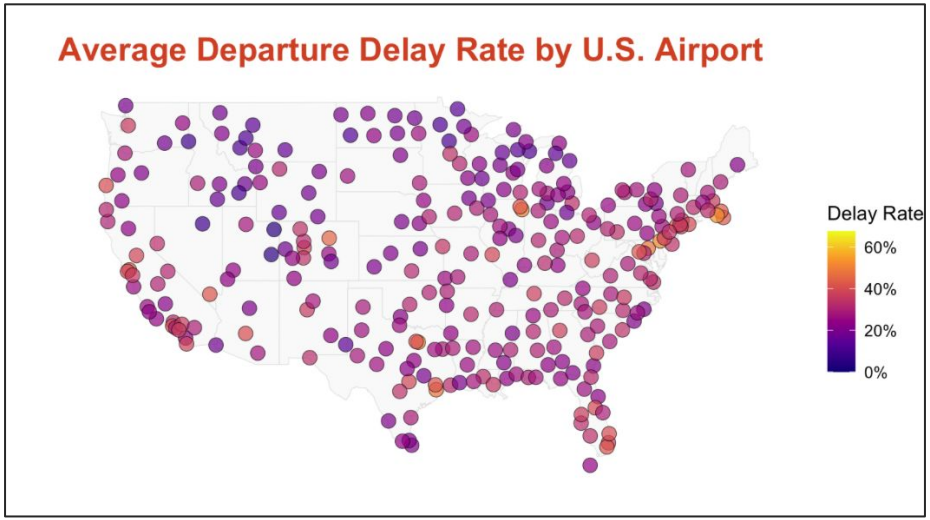
Understanding When, Where, and Why Delays Happen

Exploratory Analysis: Correlation Heatmap



- What factors actually predict delays?
  - **High Positive Correlations**
    - Departure Delay & Arrival Delay → 0.96
    - Distance & Scheduled Time → 0.98
    - Airline Delay & Departure Delay → 0.66
    - Late Aircraft Delay & Departure Delay → 0.63
  - **Low/Moderate Correlations**
    - Air System Delay & Departure Delay → 0.28
    - Weather Delay & Departure Delay → 0.26
- The strongest correlations guided predictor selections for our models

Exploratory Analysis: Where are the worst delays?

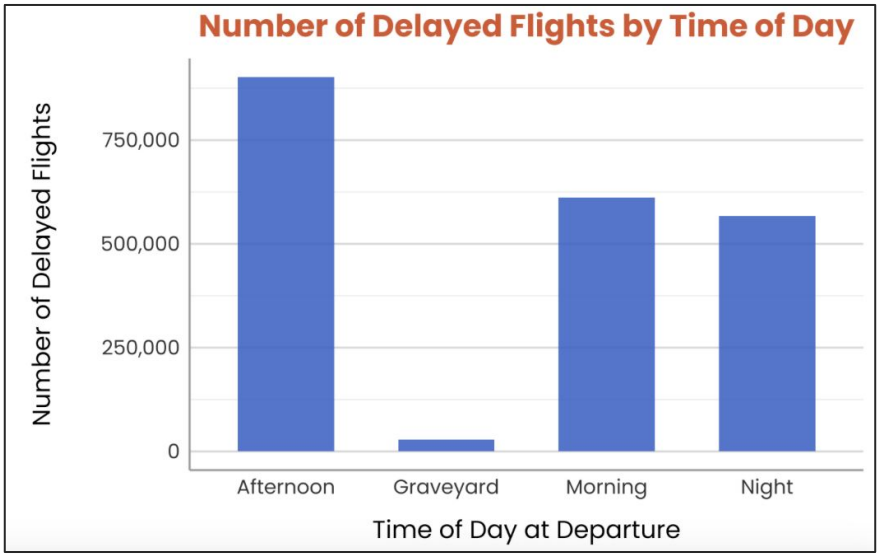


Regional hotspots: Northeast, California, and Southeast airports show the highest delay rates.

Smaller airports not immune to delays: Some midwest hubs also face frequent delays.

Operational takeaway: Delay-prone regions may benefit from targeted resource allocation and scheduling adjustments.

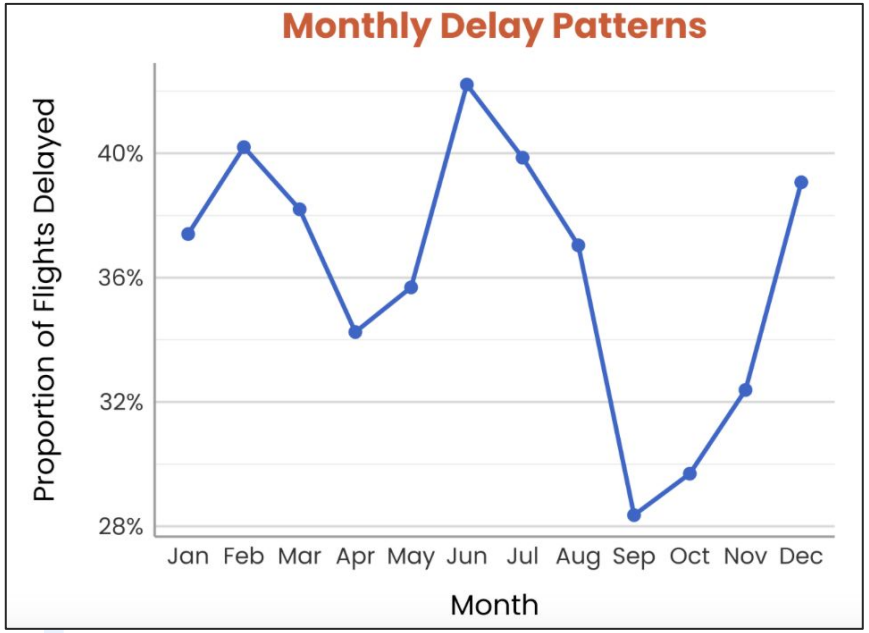
Exploratory Analysis: When are delays most likely?



Operational “Snowball Effect”

- Afternoon flights face peak delays due to cumulative disruptions
- Strategic scheduling & recovery planning are key to minimizing afternoon impact

Exploratory Analysis:
How do delays vary throughout the year?



- Summer peak driven by heavy travel and weather disruptions such as thunderstorms
- Sharp decline in autumn months likely due to post-summer travel lull in addition to stable weather conditions
- Year-end increases reflect holiday travel surges and winter storms

3

Regression Modeling

Bivariate and Multivariate Models

Bivariate Modeling: Simple Linear Model

Variable	Cor_Departure_Delay
DEPARTURE_DELAY	1.00000000
ARRIVAL_DELAY	0.96318133
AIRLINE_DELAY	0.65516525

Output Variable: DEPARTURE_DELAY

Input Variable: AIRLINE_DELAY

```
Call:
lm(formula = DEPARTURE_DELAY ~ AIRLINE_DELAY, data = flights_train)

Residuals:
    Min      1Q  Median      3Q     Max
-103.75  -18.92  -12.86    0.10 1437.08

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.992e+01  2.996e-02     665   <2e-16 ***
AIRLINE_DELAY 9.982e-01  8.801e-04    1134   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38.01 on 1710441 degrees of freedom
Multiple R-squared:  0.4292,    Adjusted R-squared:  0.4292
F-statistic: 1.286e+06 on 1 and 1710441 DF,  p-value: < 2.2e-16
```

(Intercept)

When there is no airline delay, departure delay is expected to be ~20 minutes on average

(AIRLINE_DELAY)

1 minute increase in airline delay is associated with a ~1 minute increase in departure delay

R-Squared

Airline_Delay explains 42.92% of the variation in departure delay

Bivariate Modeling: RIDGE Regression

Unregularized model: DEPARTURE_DELAY ~ AIRLINE_DELAY + AIRLINE_DELAY^2 + AIRLINE_DELAY^3 + AIRLINE_DELAY^4

REGULARIZING the 4th order polynomial model leveraging lmridge()

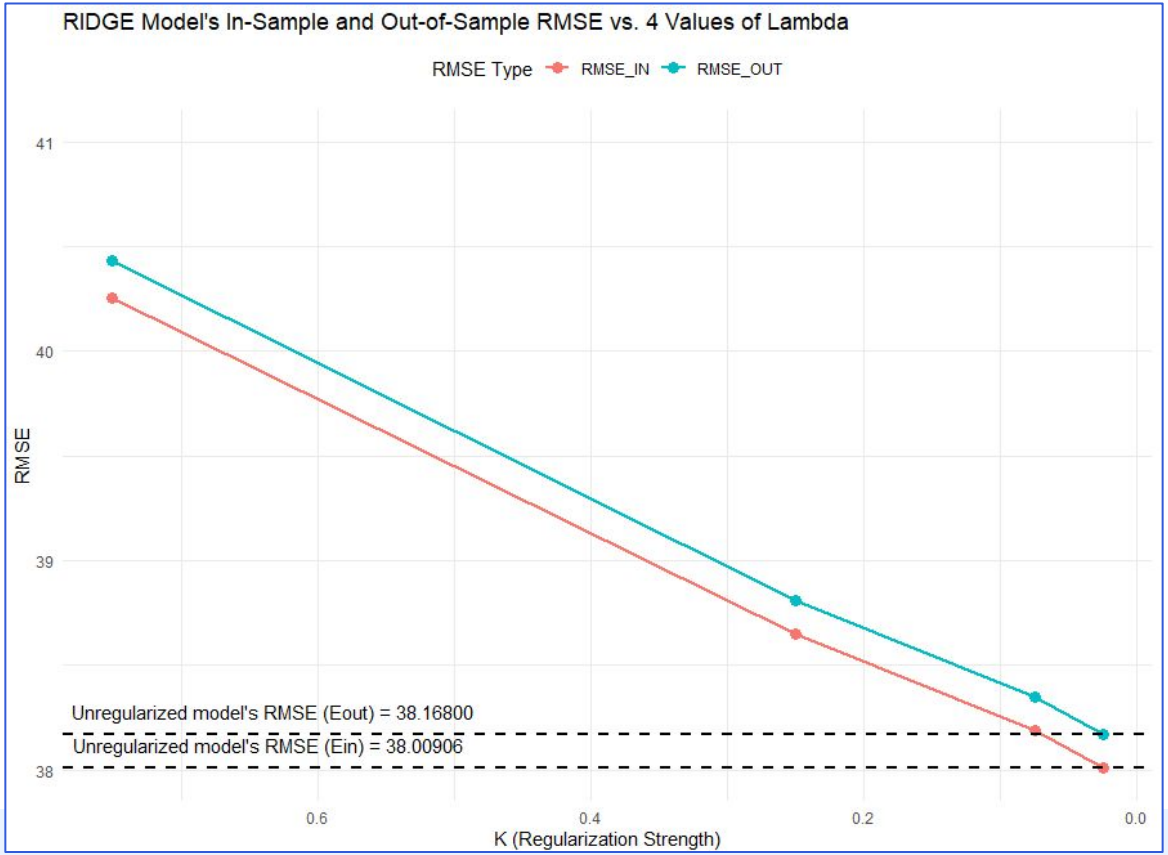
WE LEARNED THAT testing many lambda values was computationally expensive

WHAT WE DID testing a few lambda values, one time

```
> TABLE_VAL_RIDGE
              K=0.75   K=0.25  K=0.075  K=0.025     UNREG
RMSE_IN   40.25567 38.64745 38.18686 38.06209 38.00906
RMSE_OUT 40.42987 38.81153 38.34554 38.22019 38.16800
```

Observe: When cost function penalty decreases, so does RMSE

Visualizing the Effect of Regularization



Bivariate Modeling: SPLINE Regression

WHY SPLINE? **Minutes of delay variable is a continuous, numeric output**

```
Approximate significance of smooth terms:
                    edf Ref.df      F p-value
s(AIRLINE_DELAY) 8.685   8.96 143644  <2e-16 ***
```

```
R-sq.(adj) =  0.429   Deviance explained = 42.9%
GCV = 1444.6  Scale est. = 1444.6     n = 1710443
```

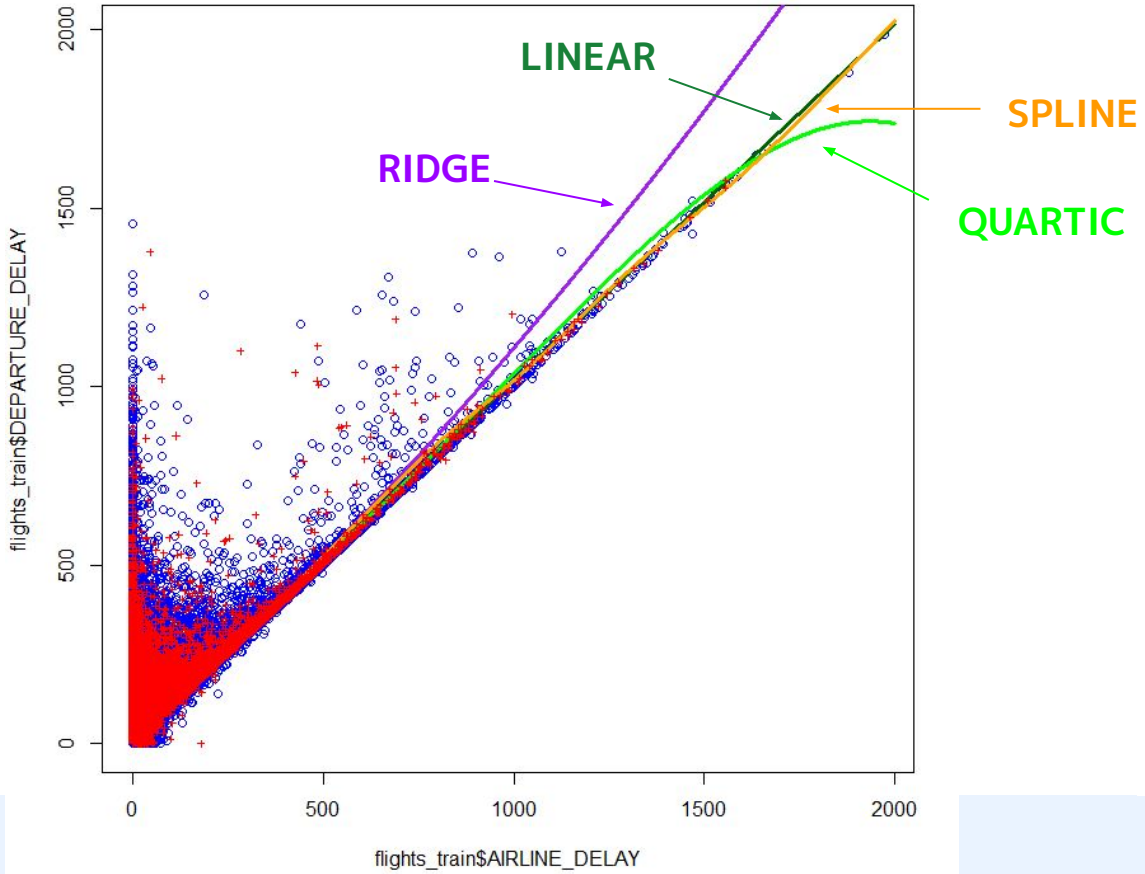
- Spline term is using most of the flexibility allowed
- High F-statistic and low p-value indicate the non-linear relationship captured is statistically significant

Benchmarking SPLINE Model Performance Against All Bivariate Models

	LINEAR	QUARTIC	RIDGE	SPLINE
RMSE_IN	38.01200	38.00906	38.06209	38.00713
RMSE_OUT	38.01200	38.16800	38.22019	38.16498

Best in-sample performance: SPLINE regression model
Best out-of-sample performance: LINEAR regression model

Visualizing Model Performances



Out of Sample Error of Bivariate Linear Model

- The linear model predicting minutes of departure delay was the best performing model on the validation set, it had the lowest RMSE

```
> TABLE_CHOSEN
            EIN  E[EOUT]     EOUT
RMSE 38.01200 38.16955 38.46936
```

- It can be seen that the model is generalizing well because its performance remained consistent on unseen data
- The model's predictions for minutes of departure delay are off by approximately ~38 minutes on average

Multivariate Regression: Linear Model

Variable	Cor_Departure_Delay
DEPARTURE_DELAY	1.00000000
ARRIVAL_DELAY	0.96318133
AIRLINE_DELAY	0.65516525
LATE_AIRCRAFT_DELAY	0.62648670
AIR_SYSTEM_DELAY	0.28217188
WEATHER_DELAY	0.26129387

Target Variable:
DEPARTURE_DELAY

Regressors:
AIRLINE_DELAY + LATE_AIRCRAFT_DELAY + AIR_SYSTEM_DELAY + WEATHER_DELAY

```
Call:
lm(formula = DEPARTURE_DELAY ~ AIRLINE_DELAY + LATE_AIRCRAFT_DELAY +
    WEATHER_DELAY + AIR_SYSTEM_DELAY, data = flights_train)

Residuals:
    Min      1Q  Median      3Q     Max
-185.09   -5.29   -1.29    5.71  582.72

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.2887165  0.0098706   535.8   <2e-16 ***
AIRLINE_DELAY     0.9932436  0.0002686  3697.9   <2e-16 ***
LATE_AIRCRAFT_DELAY 0.9978065  0.0002891  3450.9   <2e-16 ***
WEATHER_DELAY     0.9356422  0.0006605  1416.7   <2e-16 ***
AIR_SYSTEM_DELAY  0.6844440  0.0004655  1470.2   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.6 on 1710438 degrees of freedom
Multiple R-squared:  0.9469,    Adjusted R-squared:  0.9469
F-statistic: 7.62e+06 on 4 and 1710438 DF,  p-value: < 2.2e-16
```

R-Squared

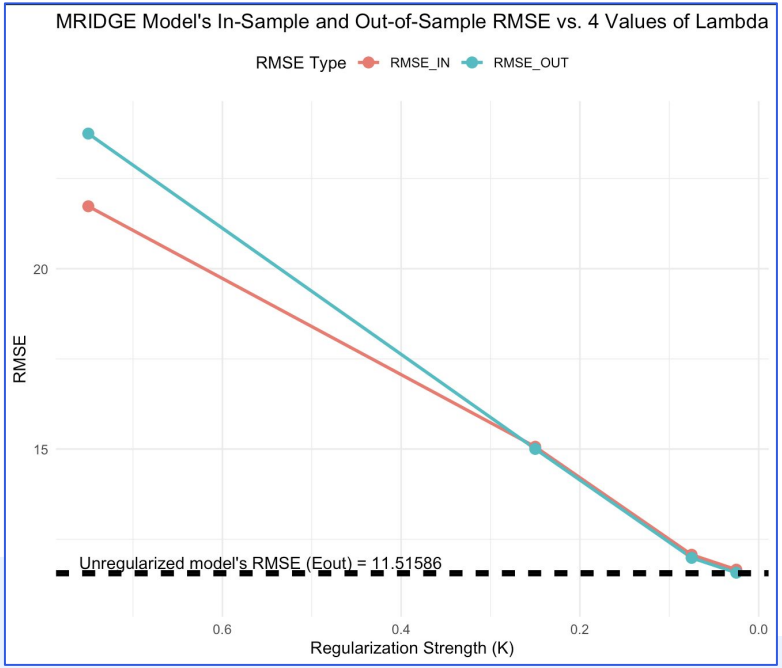
- 94.69% of the variability in minutes of departure delay captured.
- 51.77% increase in explanatory power from the bivariate model
- Adjusted R^2 is the same due to the number of observations in the dataset

Significance

- All estimated coefficients are significant***
- Large F-statistic indicates that model has overall significance

Regularizing the Multivariate Regression Model

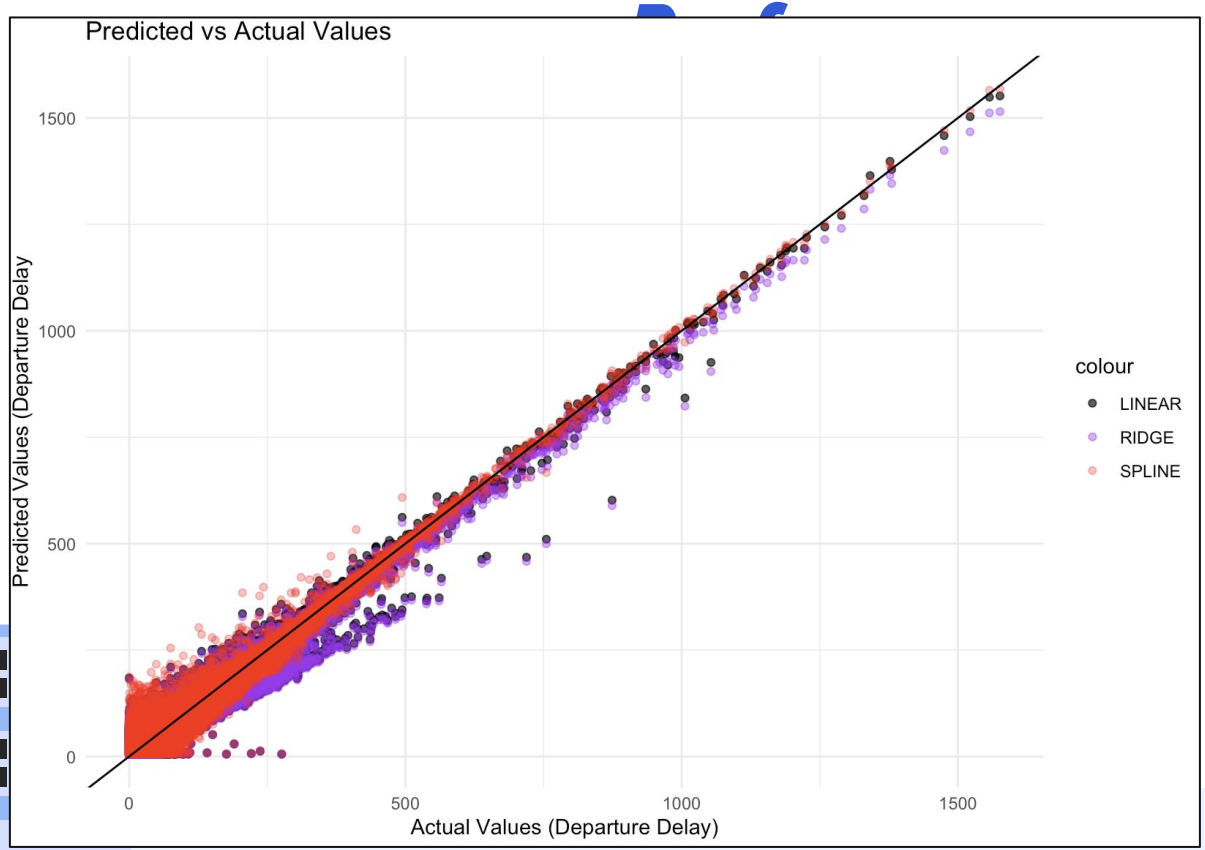
```
> TABLE_VAL_MM_RIDGE
          K=0.75   K=0.25  K=0.075  K=0.025    UNREG
RMSE_IN  23.73083 15.06338 12.07047 11.65685 11.59823
RMSE_OUT 23.74574 15.00398 11.98404 11.57112 11.51586
```



Observations:

- As **regularization decreases**, in-sample and estimated out-of-sample **errors decrease**
- When the model is **overly-penalized** by lambda, it begins to **underfit**, leading to **greater errors** on average
- The RIDGE models' errors begin to converge toward the unregularized model's prediction error of 11. 6 minutes of delay on average

Comparing Multivariate Model



It can be seen *that the SPLINE regression model is making predictions closest to the actual values of departure delay minutes.*

Multivariate Support Vector Regression

The Problem

- SVM creates an n x n matrix
- 1.7 million training observations meant computer would have to store a matrix with almost 3 trillion distance metrics
- In short, computations exceeded capacity of computers and would take days to run

The Solution: Principal Component Analysis

- PCA is a unsupervised learning technique for reducing dimensionality of features
- This brought our features down into a 2-dimensional space, making convergence speeds very rapid

Principal Components

```
> summary(flights_train.pc1)
Importance of components:
                          PC1     PC2     PC3     PC4
Standard deviation     11.7991 10.7882 10.0412 3.90671
Proportion of Variance  0.3745  0.3131  0.2713 0.04106
Cumulative Proportion   0.3745  0.6877  0.9589 1.00000
```

```
> summary(flights_valid.pc1)
Importance of components:
                          PC1     PC2     PC3     PC4
Standard deviation     11.8351 10.7203 10.0295 3.93553
Proportion of Variance  0.3775  0.3097  0.2711 0.04174
Cumulative Proportion   0.3775  0.6872  0.9583 1.00000
```

← Lilie's R screen time this week 😳

Multivariate Support Vector Regression

```
Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  linear
       cost:  1
      gamma:  0.25
    epsilon:  0.1


Number of Support Vectors:  19495
```

Subsetting the data:
Even with our data in lower dimensional space, we still had to use a subset of the training set for computational efficiency

- Training: 20,000 reshuffled rows
- Testing: 2,000 reshuffled rows

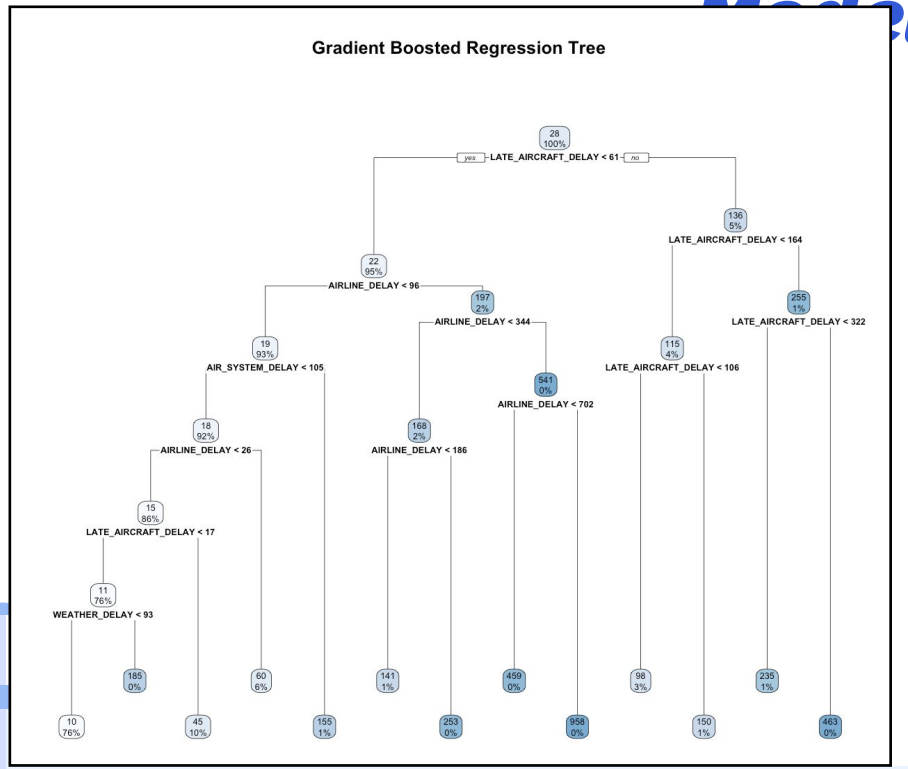
19495 support vectors:
The number of support vectors is very close to the number of observations the model was trained on

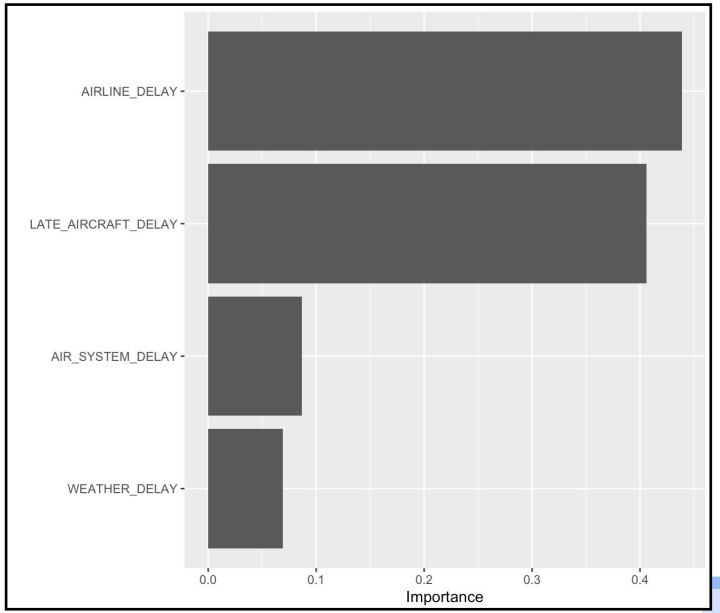
- Risk of overfitting

RMSE_IN	RMSE_OUT
12.56635	13.43854

The estimated out of sample error is slightly higher than the in-sample error. This means the SVM model is generalizing well and that its estimated to predict incorrectly by 13.44 minutes on average, on unseen data.

Gradient Boosted Regression Tree Model





Multivariate Model Selection

	LINEAR	RIDGE	SPLINE	TUNED	GXBOOST
RMSE IN	11.598228	11.656852	10.051247	8.937537	9.030226
RMSE OUT	11.515862	11.571123	10.002503	9.518980	9.205486

Gradient Boosted Out-of-Sample Error

```
# A tibble: 1 × 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard        9.31
```

Best in-sample performance

- Tuned regression tree with a cost parameter of 0.0000000001

Best estimated out-of-sample performance

- Gradient boosted regression tree

Out of sample error: The gradient boosted regression tree incorrectly predicts minutes of departure delay by 9.31 minutes on average

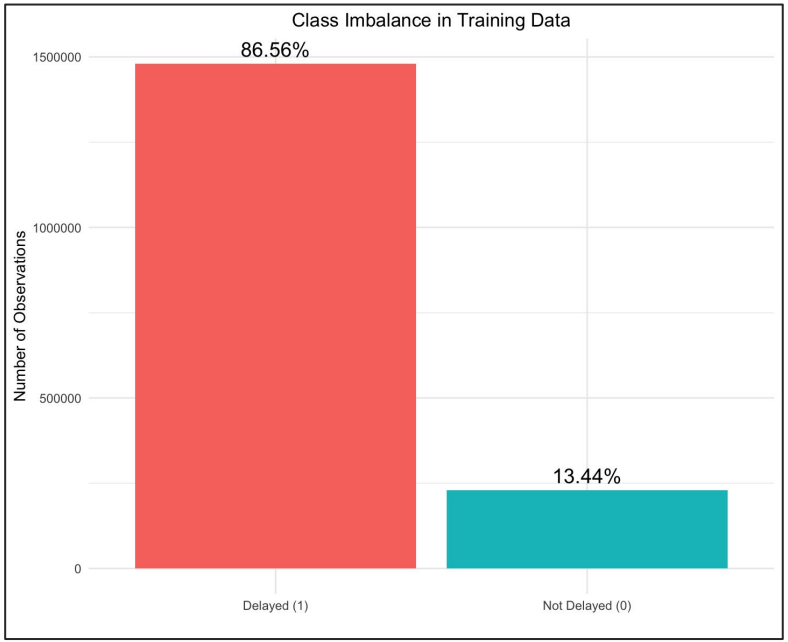
4

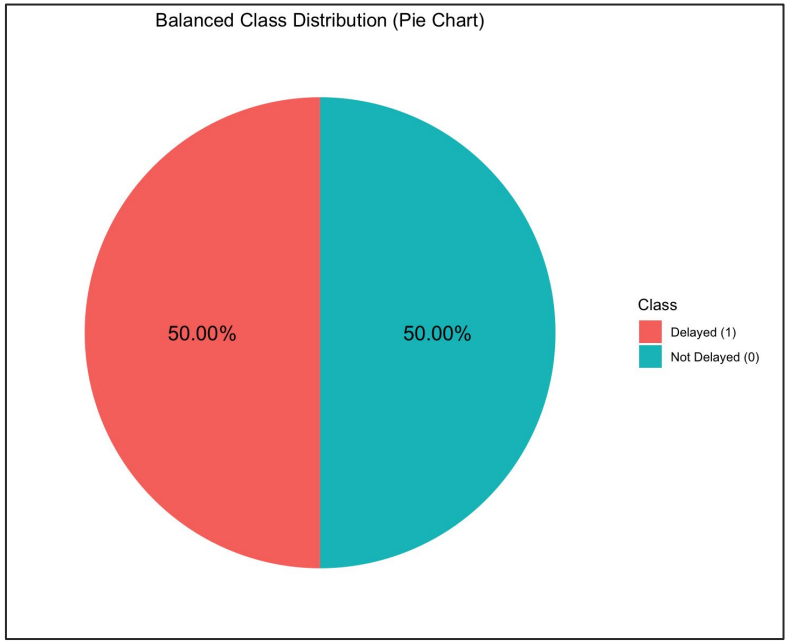
Classification Modeling

Binary and Multiclass Approaches

Binary Tools

Balanced Output





Logit Model: Predictors & Results

```
Call:
glm(formula = DELAYED ~ MONDAY + TUESDAY + THURSDAY + FRIDAY +
    SCHEDULED_GRAVEYARD_DEPARTURE + SCHEDULED_MORNING_DEPARTURE +
    SCHEDULED_AFTERNOON_DEPARTURE + CAT_DELAY_RANK, family = "binomial",
    data = flights_train_bal)
```

```
Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                    0.701991   0.015833  44.338  < 2e-16 ***
MONDAY                         0.063777   0.009022   7.069 1.56e-12 ***
TUESDAY                        0.028861   0.009297   3.104  0.00191 **
THURSDAY                       0.059980   0.008933   6.715 1.89e-11 ***
FRIDAY                         0.036432   0.008948   4.072 4.67e-05 ***
SCHEDULED_GRAVEYARD_DEPARTURE -1.101114   0.023035 -47.801  < 2e-16 ***
SCHEDULED_MORNING_DEPARTURE   -0.729247   0.008016 -90.976  < 2e-16 ***
SCHEDULED_AFTERNOON_DEPARTURE -0.219917   0.007790 -28.230  < 2e-16 ***
CAT_DELAY_RANK2               -0.386695   0.015128 -25.561  < 2e-16 ***
CAT_DELAY_RANK3               -0.379942   0.015167 -25.051  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 637554  on 459897  degrees of freedom
Residual deviance: 625921  on 459888  degrees of freedom
AIC: 625941

Number of Fisher Scoring iterations: 4
```

		2.5 %	97.5 %
(Intercept)	2.0177659	1.9561583	2.0814107
MONDAY	1.0658544	1.0471734	1.0848699
TUESDAY	1.0292818	1.0106960	1.0482096
THURSDAY	1.0618157	1.0433878	1.0805703
FRIDAY	1.0371038	1.0190744	1.0554526
SCHEDULED_GRAVEYARD_DEPARTURE	0.3325006	0.3177893	0.3478213
SCHEDULED_MORNING_DEPARTURE	0.4822719	0.4747525	0.4899067
SCHEDULED_AFTERNOON_DEPARTURE	0.8025855	0.7904222	0.8149321
CAT_DELAY_RANK2	0.6792983	0.6594379	0.6997269
CAT_DELAY_RANK3	0.6839012	0.6638564	0.7045213

- All factors hold strong statistical significance
- We see that earlier morning flights are less likely to be delayed
- Delays are slightly more common on weekdays specially thursdays
- This model was trained on balanced data, helping it avoid bias toward predicting the majority class (delayed)

Probit Model: Same Predictors, Same Results

```
Call:
glm(formula = DELAYED ~ MONDAY + TUESDAY + THURSDAY + FRIDAY +
    SCHEDULED_GRAVEYARD_DEPARTURE + SCHEDULED_MORNING_DEPARTURE +
    SCHEDULED_AFTERNOON_DEPARTURE + CAT_DELAY_RANK, family = binomial("probit"),
    data = flights_train_bal)
```

```
Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                    0.437491   0.009801  44.638  < 2e-16 ***
MONDAY                         0.039725   0.005623   7.065 1.61e-12 ***
TUESDAY                        0.017940   0.005795   3.096  0.00196 **
THURSDAY                       0.037268   0.005568   6.694 2.18e-11 ***
FRIDAY                         0.022575   0.005577   4.048 5.17e-05 ***
SCHEDULED_GRAVEYARD_DEPARTURE -0.685289   0.014055 -48.758  < 2e-16 ***
SCHEDULED_MORNING_DEPARTURE   -0.455729   0.004986 -91.403  < 2e-16 ***
SCHEDULED_AFTERNOON_DEPARTURE -0.137258   0.004857 -28.262  < 2e-16 ***
CAT_DELAY_RANK2               -0.240458   0.009372 -25.658  < 2e-16 ***
CAT_DELAY_RANK3               -0.235979   0.009396 -25.115  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 637554  on 459897  degrees of freedom
Residual deviance: 625921  on 459888  degrees of freedom
AIC: 625941

Number of Fisher Scoring iterations: 4
```

- Same predictors as the Logit model
- Coefficients and signs are consistent
- All key variables remain statistically significant
- Model fit metrics (AIC, deviance) are nearly identical
- Logit and Probit produce equivalent insights

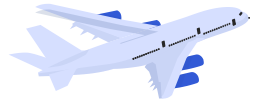
Metrics

```
Confusion Matrix and Statistics

          Reference
Prediction      0      1
         0  21461  92666
         1  27789 224608

               Accuracy : 0.6714
                 95% CI : (0.6698, 0.6729)
    No Information Rate : 0.8656
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0923

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7079
            Specificity : 0.4358
         Pos Pred Value : 0.8899
         Neg Pred Value : 0.1880
             Prevalence : 0.8656
         Detection Rate : 0.6128
   Detection Prevalence : 0.6886
      Balanced Accuracy : 0.5718

       'Positive' Class : 1
```

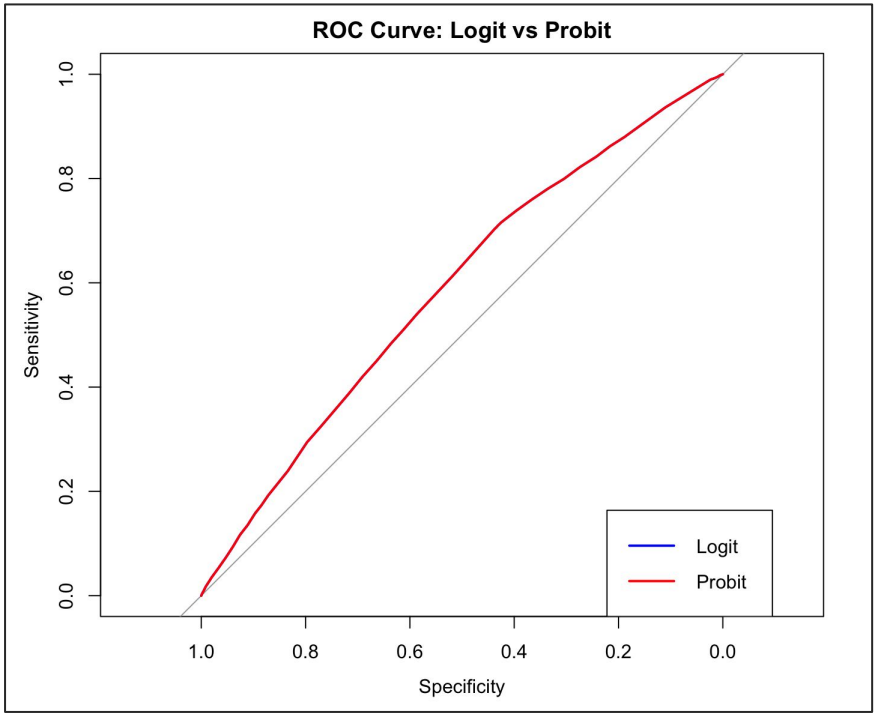
Logit

```
Confusion Matrix and Statistics

          Reference
Prediction      0      1
         0  21555  93348
         1  27695 223926

               Accuracy : 0.6698
                 95% CI : (0.6682, 0.6713)
    No Information Rate : 0.8656
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0918

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7058
            Specificity : 0.4377
         Pos Pred Value : 0.8899
         Neg Pred Value : 0.1876
             Prevalence : 0.8656
         Detection Rate : 0.6109
   Detection Prevalence : 0.6865
      Balanced Accuracy : 0.5717

       'Positive' Class : 1
```

Probit

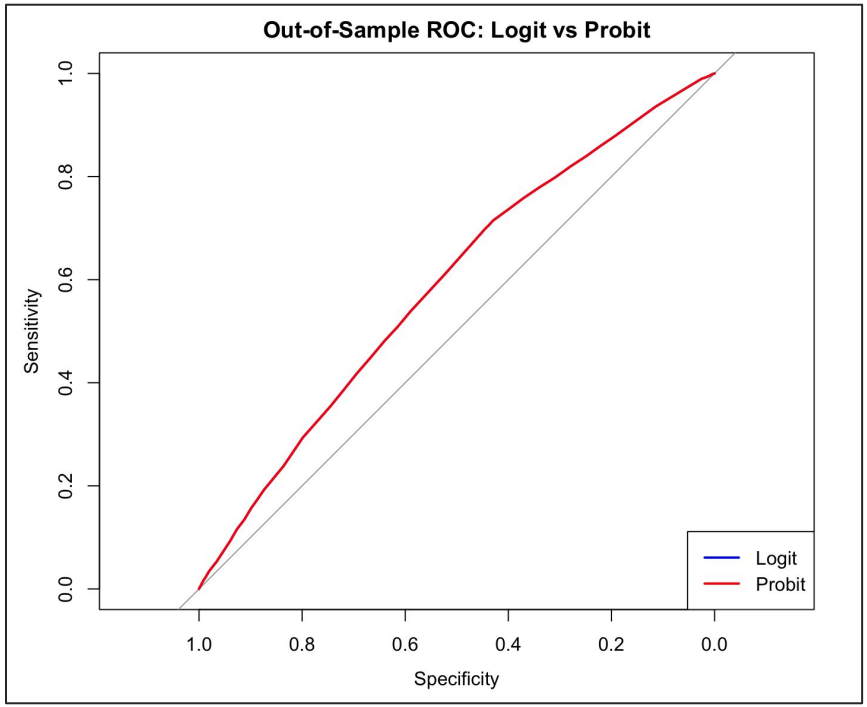


ROC Curve: Logit vs. Probit



- ROC curve comparison between Logit and Probit models
- Both models perform slightly better than random guessing
- **AUC (Logit): 0.59 | AUC (Probit): 0.58**
- Logit slightly outperforms Probit in classification
- These results show statistical significance but limited real-world predictability

Out-of-Sample Performance



- Both models have similar performance
- Logit slightly outperforms Probit in accuracy
- Out-of-sample errors indicate moderate generalization
- Results show room for improvement with more complex models

Naive Model Comparison

Do We Beat a Naive Guess?

- AUC of naive classifier = 0.5
- Models AUC = 0.5869
- Emphasize that models outperform random guessing, even if only modestly

Model Type	In-Sample	Out-Sample	Accuracy
Logit	0.4295	0.3286	67.14%
Probit	0.4296	0.3302	66.98%

Comparing Four Classification Approaches

SVM

Binary model to predict On-Time vs. Delayed flights

CART

Simple multiclass model to predict delay severity

Random Forest

Ensemble model improving multiclass prediction accuracy

XGBoost

Boosted trees to refine predictions and handle class imbalance

Binary Model: Support Vector Machine

- **Feature Engineering ➔** One-hot encoded Time of Day (morning, afternoon, night, graveyard) and Day of Week. Also, created Airline Delay Rank based on carrier performance.
- **Sampling for Feasibility ➔** Pulled 20,000 observations for training and 2,000 for validation to avoid computational overload.
- **Best Performing Model ➔ 13.4% out-of-sample error.** High accuracy driven by simple binary task (delay vs. on-time). Limited in predicting delay severity, leading us to explore other models.

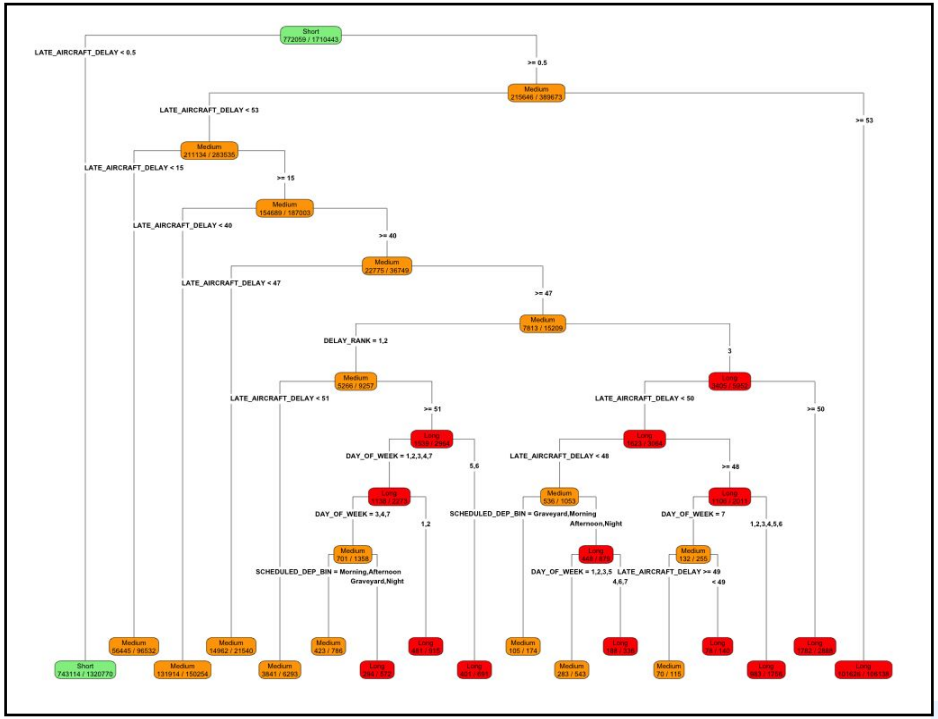

```
SVM_Model <- svm(DEPARTURE_DELAY ~
                   SCHEDULED_MORNING_DEPARTURE +
                   SCHEDULED_AFTERNOON_DEPARTURE +
                   SCHEDULED_NIGHT_DEPARTURE +
                   MONDAY + TUESDAY + WEDNESDAY + THURSDAY +
                   FRIDAY + SATURDAY + SUNDAY +
                   LATE_AIRCRAFT_DELAY + DELAY_RANK,
                 data = flights_train_small,
                 type = "C-classification",
                 kernel = "radial",
                 cost = 1,
                 gamma = 1 / (ncol(flights_train_small) - 1),
                 scale = FALSE)
```

Multiclass Model: Classification Tree

Before Pruning

```
cart_spec <- decision_tree(
  min_n = 5,
  tree_depth = 30,
  cost_complexity = 0.00001
) %>%
  set_engine("rpart") %>%
  set_mode("classification")

# Updated formula
cart_tree <- cart_spec %>%
  fit(DEPARTURE_DELAY_CAT ~
        DAY_OF_WEEK +
        SCHEDULED_DEP_BIN +
        LATE_AIRCRAFT_DELAY +
        DELAY_RANK,
      data = flights_train)
```

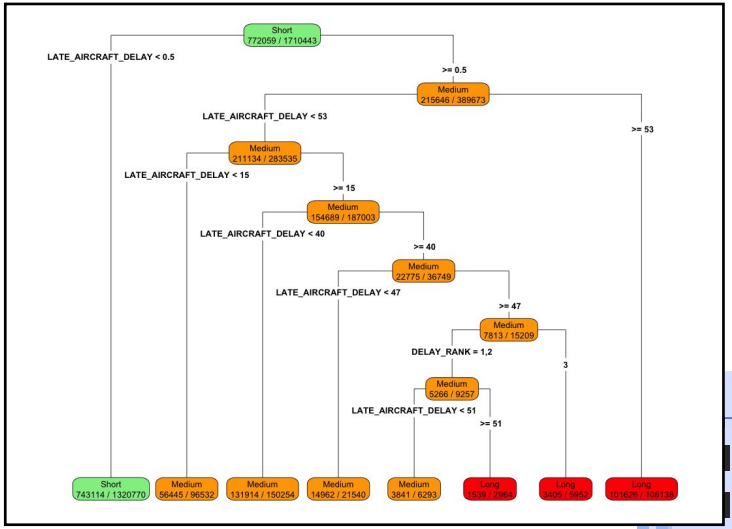
```
rpart.plot(cart_tree$fit,
           type = 4,
           extra = 2,
           roundint = FALSE,
           box.col = custom_colors[cart_tree$fit$frame$yval])
```



Multiclass Model: Classification Tree

After Pruning

```
# View cp table to confirm
printcp(cart_tree$fit)

# Get optimal cp
optimal_cp <- printcp(cart_tree$fit)[which.min(printcp(cart_tree$fit)[,"xerror"]), "CP"]

# Now explicitly call prune from rpart
pruned_tree <- rpart::prune(cart_tree$fit, cp = optimal_cp)

# Plot pruned tree
# Exact custom colors per class level
custom_colors <- c("green", "lightgreen", "orange", "red")

rpart.plot(pruned_tree,
           type = 4,
           extra = 2,
           roundint = FALSE,
           box.col = custom_colors[pruned_tree$frame$yval])
```



Validation Accuracy ➔ 61.9%
No Information Rate ➔ 45.2%

Ensemble Model: Random Forest

```
Confusion Matrix and Statistics


          OnTime  Short Medium   Long
  OnTime       0      0      0      0
  Short    49250 159494  57657  16764
  Medium       0   6220  44495   8175
  Long         0      1   1693  22775

Overall Statistics

               Accuracy : 0.6187
                 95% CI : (0.6171, 0.6203)
    No Information Rate : 0.4521
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3607

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: OnTime Class: Short Class: Medium Class: Long
Sensitivity                 0.0000       0.9625        0.4285     0.47732
Specificity                 1.0000       0.3841        0.9452     0.99469
Pos Pred Value                 NaN       0.5633        0.7556     0.93077
Neg Pred Value              0.8656       0.9254        0.8071     0.92709
Prevalence                  0.1344       0.4521        0.2833     0.13018
Detection Rate              0.0000       0.4352        0.1214     0.06214
Detection Prevalence        0.0000       0.7726        0.1607     0.06676
Balanced Accuracy           0.5000       0.6733        0.6868     0.73600
```

- **Model Overview**
  - 100 trees, 4 predictors (Late Aircraft Delay, Airline Rank, Day of Week, Departure Time)
- **Handles complexity better**
  - Boosted sensitivity for Medium (42.8%) & Long delays (47.7%)
- **OnTime flights still missed**
  - 0% sensitivity shows class imbalance issue persists
- **Stable generalization**
  - Out-of-sample error = **38.1%**, mirrors CART but with stronger minority class detection

Ensemble Model: XGBoost

```
Confusion Matrix and Statistics


          OnTime  Short Medium   Long
  OnTime       0      0      0      0
  Short    49250 159423  57549  16749
  Medium       0   6292  44777   8307
  Long         0      0   1519  22658

Overall Statistics

               Accuracy : 0.6189
                 95% CI : (0.6174, 0.6205)
    No Information Rate : 0.4521
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3611

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: OnTime Class: Short Class: Medium Class: Long
Sensitivity                 0.0000       0.9620        0.4312     0.47487
Specificity                 1.0000       0.3847        0.9444     0.99524
Pos Pred Value                 NaN       0.5634        0.7541     0.93717
Neg Pred Value              0.8656       0.9247        0.8077     0.92681
Prevalence                  0.1344       0.4521        0.2833     0.13018
Detection Rate              0.0000       0.4350        0.1222     0.06182
Detection Prevalence        0.0000       0.7720        0.1620     0.06596
Balanced Accuracy           0.5000       0.6734        0.6878     0.73505
```

- **Model Overview**
  - Boosted trees using the same predictors as previous RF model
- **Performance**
  - Validation accuracy & error rate (E_OUT = 38.1%) align with Random Forest
- **Moderate Gains**
  - Slightly stronger in Medium & Long delays, but no improvement for OnTime flights
- **Class Imbalance**
  - Still skewed toward delay-heavy predictions
  - Dominant variables overpower minority classes.
- *Note on Rebalancing: More balanced data models explored in report (excluded here for focus & consistency)*

5

Summarized Results

Accuracy and Error Statistics

Regression Model Comparison

Model Type	In-Sample Error	Estimated Out-of-Sample Error	Out of Sample Error
Bivariate Linear Regression Model	38.012	38.17	38.47
Gradient Boosted Regression Tree	9.03	9.21	9.31

Error metric, RMSE: number of minutes of departure delay that predictions are off by, on average

Classification Model Comparison

Model Type	In-Sample Error	Out-of-Sample Error
Logistic Model (Logit)	0.4295	0.3286
Probability Unit Model (Probit)	0.4296	0.3302
SVM Binary Classification (Unbalanced)	0.1344	0.1344
CART Multiclass Classification (Unbalanced)	0.3817	0.3813
Random Forest Multiclass Classification (Unbalanced)	0.3817	0.3813
XGBoost Multiclass Classification (Unbalanced)	0.3821	0.3811

6

Conclusion and Next Steps

Project limitations, business insights, and recommendations

Project Limitations

Imbalanced & Large Dataset

The dataset was extremely large and heavily imbalanced, requiring sampling and engineering to make modeling feasible.

Outdated Data

Our 2015 dataset may not reflect current flight operations due to industry changes, new technologies, and COVID-related impacts.

Broad Variable Definitions

Broad variables (e.g., “weather delay”) limited prediction accuracy; more detail would improve root cause analysis.

Business Insights & Recommendations

- **Proactive Customer Communication**
  - Binary SVM model achieved high accuracy (13.3% error) in predicting delayed flights
  - Supports real-time in-app notifications, rebooking, and proactive customer updates before disruptions escalate
- **Operational Bottlenecks & Turnaround Efficiency**
  - Late Aircraft Delay consistently emerged as the top driver of delays across models
  - Highlights the need to improve aircraft turnaround processes and implement schedule buffers to prevent cascading disruptions
- **Collaborative Delay Mitigation**
  - Many delays stem from shared airspace and airport congestion.
  - Airlines can leverage model findings to coordinate with airports and rival carriers during peak traffic periods to minimize network-wide delays
- **Following Industry Standards**
  - The Department of Transportation (DOT) requires refunds for delays over 3 hours
  - Early delay prediction enables airlines to adjust operations proactively, reducing refund liabilities and managing customer expectations
- **Resource Allocation & Prioritization**
  - Multiclass models (CART, RF, XGBoost) provided useful insights into delay severity.
  - Helps airlines prioritize gate assignments, crew rotations, and standby aircraft for more efficient disruption management

Thanks!

Let us know if you have any questions!

vbreton@sandiego.edu
lcatania@sandiego.edu
lizascott@sandiego.edu