USPS Weather Impact Challenge

Team Kangaroo

August 2, 2024

Team Kangaroo

Lowell O'Connell
Product Owner

Julie Ritenour
Scrum Master

Anderson Battin
Developer

Praveen Kumar
Developer

Jennifer Carr
Developer

Evans Sarker
Developer

Project Overview

Our Client

- The United States Postal Service (USPS), Chief Data & Analytics Office
- Our Point-of-Contact:
    - Arslan Saleem, Director, Corporate Performance Reporting & Analytics

Problem Statement

The United States Postal Service (USPS) aims to increase knowledge about how weather events impact mail and package processing, transportation and delivery. The USPS has partnered with George Mason University students to build a machine learning model that will use publicly acquired weather data to predict if a piece of mail or a package will be late.

Can weather predict if a piece of mail or a package will be late?

Picture: USPS. “Taking the bite out of frost: USPS share tips for winter weather.” Jan 3, 2024

Project Overview

Data Collection
- USPS - Mail & Package
- NOAA API – Weather

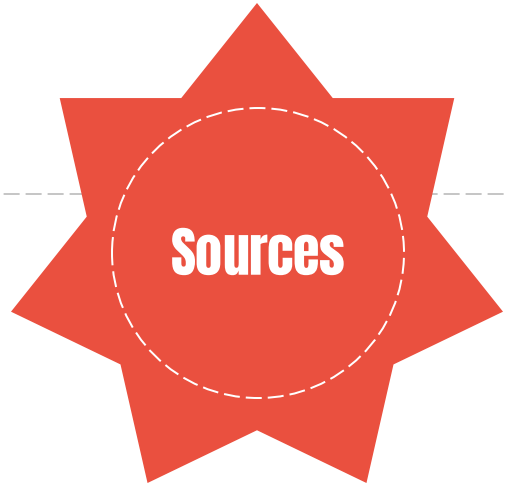
Data Preprocessing
- Data Cleaning
- Exploratory Data Analysis

Model Exploration
- Classification
- Clustering
- Neural Networks

Model Selection
- XGBoost

Project Data

Data Overview



USPS

- Provided by the customer
- Contains mail and package originating, destination and barcode scan information
- Shared via compressed file using Microsoft SharePoint

Weather

- Global Historical Climatology Network (GHCN) Daily database
- Composite of multiple weather sources
- Historical precipitation, snowfall, minimum and maximum temperatures
- Obtained via NOAA API

Data Context

- An Artic Blast occurred this year from January 12-16 and had a heavy impact on USPS operations in the Tennessee area
- USPS client team provided their data based on the Tennessee region and dates ranging from January 8-21

Data Conditioning

Cleaning

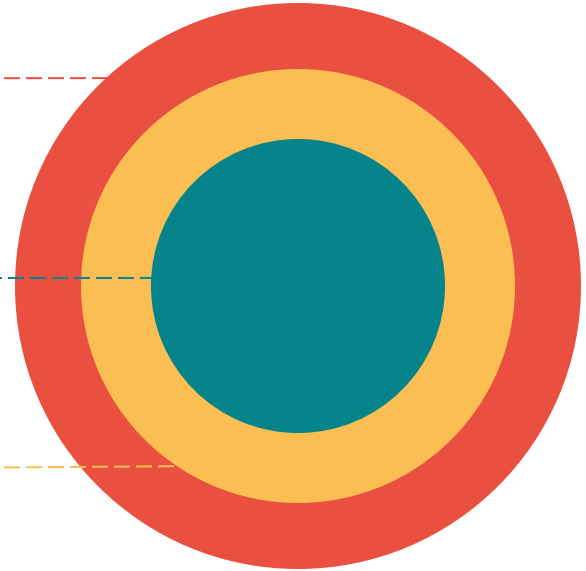
Discarded incomplete records, outliers and variables to prevent data leakage

Normalizing

Ensured consistency

Merging

Based on shared attributes such as date, geospatial location and zip code



Data Conditioning: Packages

- Raw file with 43 million records
  was parsed down to zip and event codes of interest
- Resulting 44 thousand records of aggregated Scans data was combined with the Package Origin Piece data
- Merged with weather data resulted in 17 thousand records
- A 'late' variable was created
  by checking if the Stop The Clock Date was greater than the Scheduled Delivery Date, and is represented by 1 for on-time packages or a 0 for late packages

Data Conditioning Phase Package Scans	# of Records
Raw File Package Scans	43,524,579
Only Zip3 of '370', '371', '372' And Event Code of 7, 10	49,432
Consolidation of records	44,476

Data Conditioning Phase Package Piece	# of Records
Raw File Package Piece	1,115,279
Only Zip3 of '370', '371', '372	37,290
Combining Weather and Scans	17,503

Data Conditioning: Mail

- Raw file began with 42 and 53 million records for originating and destinating mail
- Cleaned data by removing null values, duplicates and rows where the Actual Delivery Date was before the Start the Clock Date
- Resulted in 38 and 44 million records

Data Conditioning Phase Originating	# of Records
Raw File	42,174,946
Removal of Null Values	41,160,323
Removal of Duplicate IDs	41,160,140
Removal of Date Errors	38,832,134

Data Conditioning Phase Destinating	# of Records
Raw File	53,733,124
Removal of Null Values	47,058,984
Removal of Duplicate IDs	47,055,291
Removal of Date Errors	44,716,693

Data Quality

	USPS	Weather (GHCN)
Completeness	Null Values	Null Values
Consistency	Yes	Different weather stations inconsistent
Uniqueness	Yes	Yes
Integrity	Yes	Yes
Conformity	Yes	Yes
Accuracy	Invalid codes/times	Yes

Data Storage

Medium

- USPS Data: Microsoft SharePoint and PCs
- Weather Data: Github Repository

Security

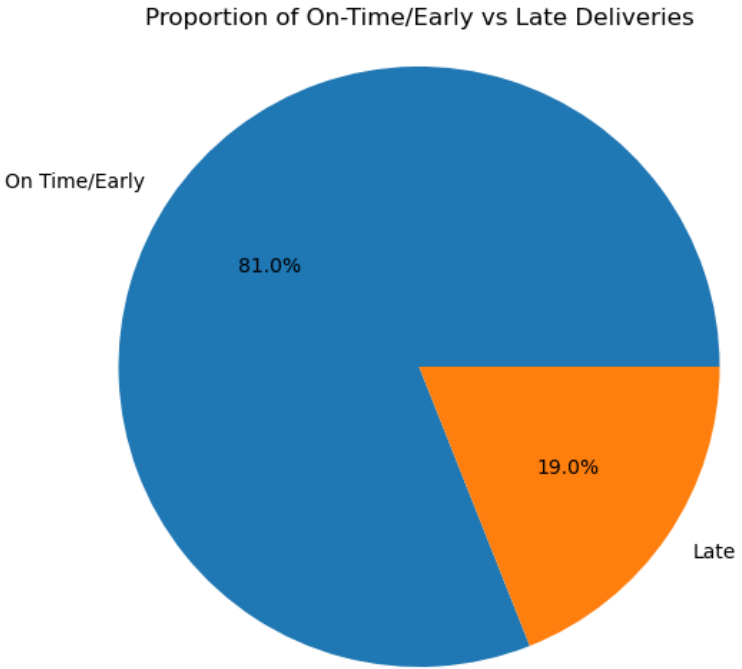
- USPS Data: Restricted to team members only
- Weather Data: Public

Costs

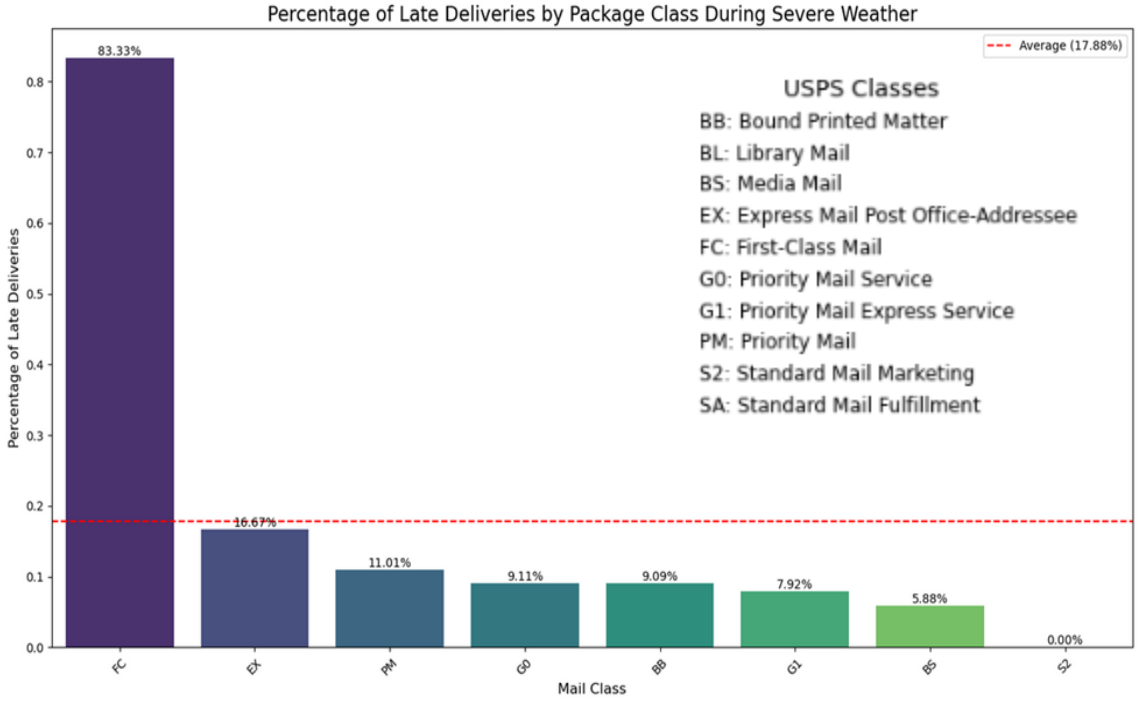
None since it is stored on PCs and GitHub is free

Exploratory Data Analysis

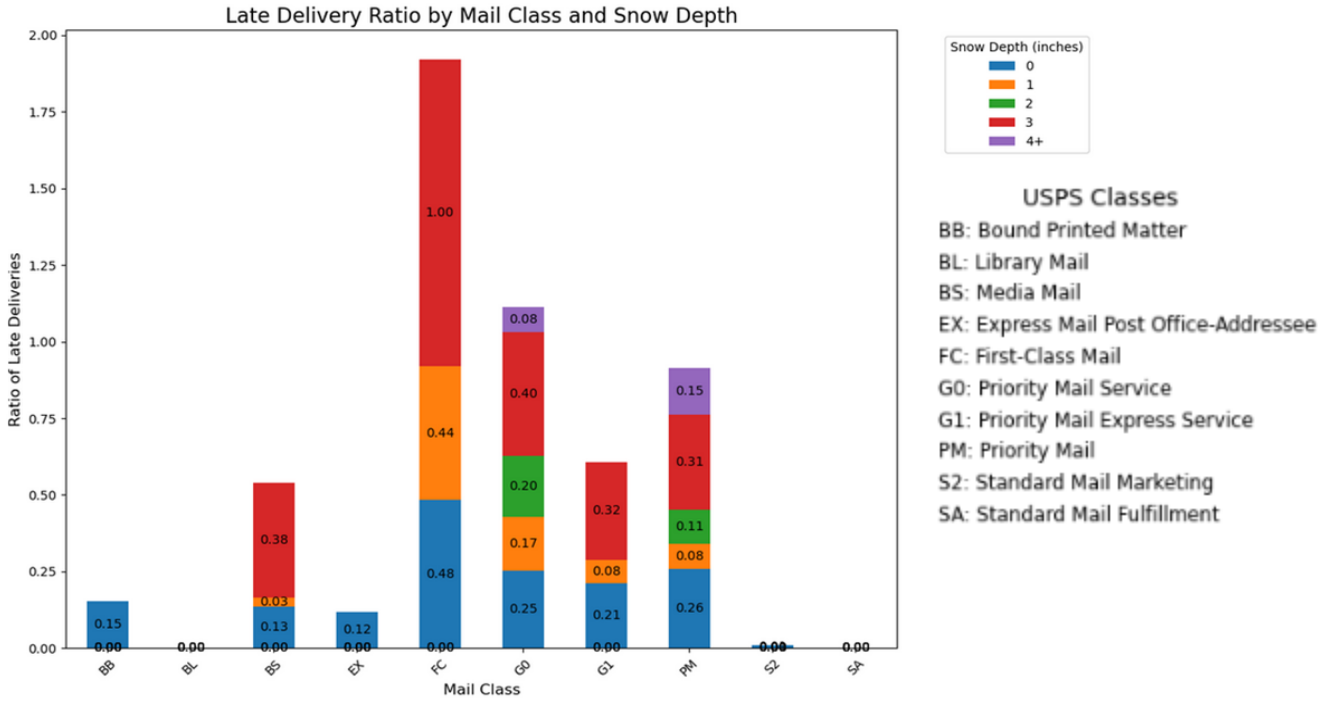
Exploratory Data Analysis- Packages



Exploratory Data Analysis- Packages



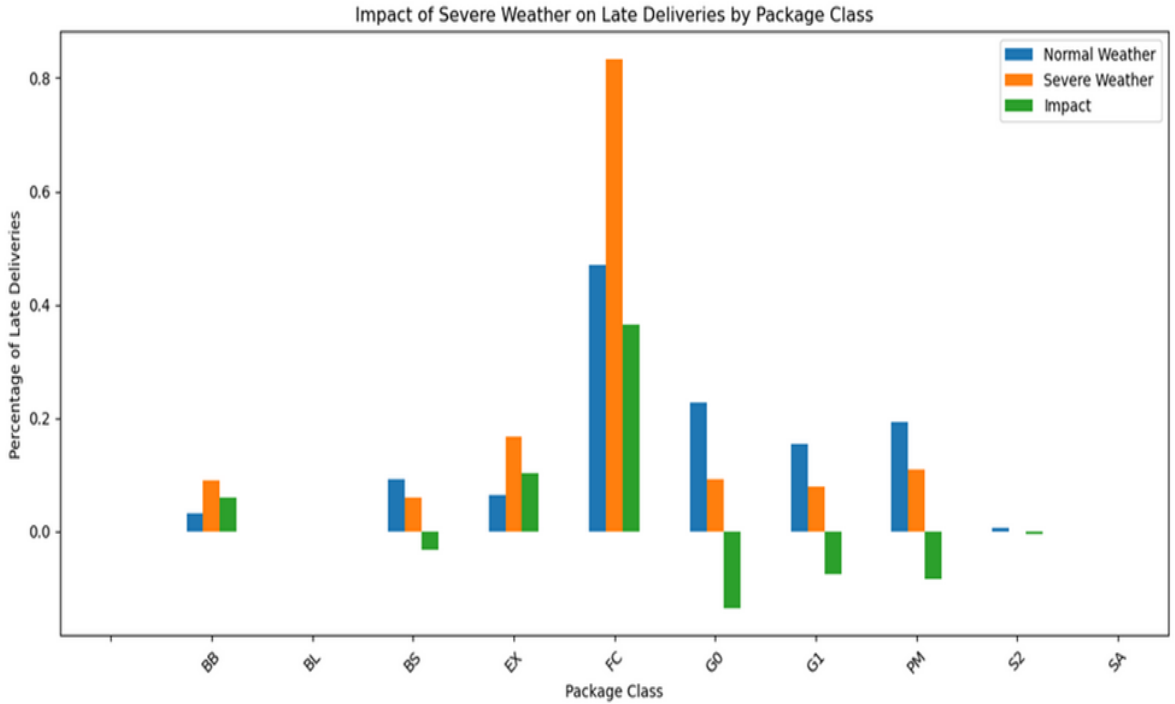
Exploratory Data Analysis- Packages



USPS Classes

BB: Bound Printed Matter
BL: Library Mail
BS: Media Mail
EX: Express Mail Post Office-Addressee
FC: First-Class Mail
G0: Priority Mail Service
G1: Priority Mail Express Service
PM: Priority Mail
S2: Standard Mail Marketing
SA: Standard Mail Fulfillment

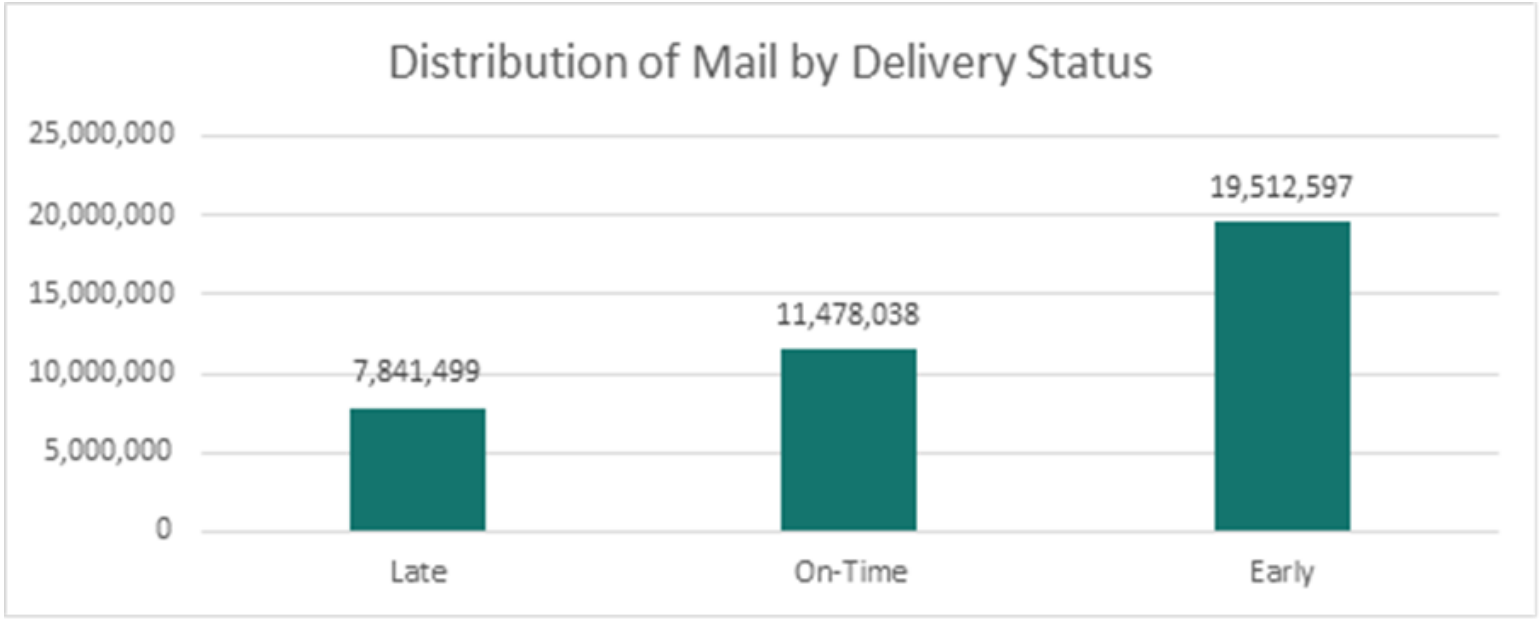
Exploratory Data Analysis- Packages



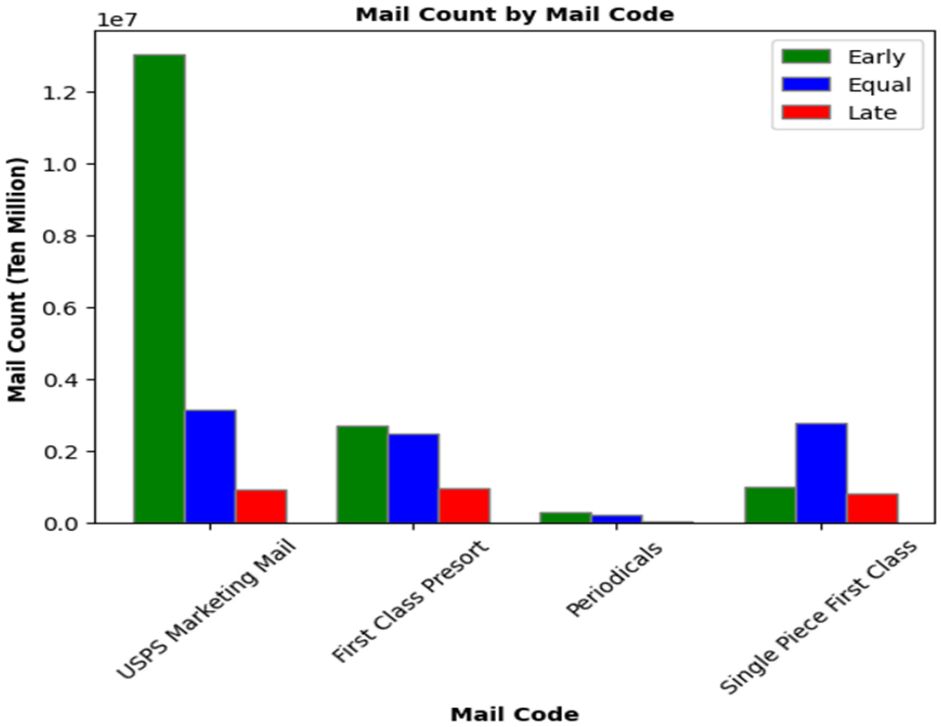
USPS Classes

BB: Bound Printed Matter
BL: Library Mail
BS: Media Mail
EX: Express Mail Post Office-Addressee
FC: First-Class Mail
G0: Priority Mail Service
G1: Priority Mail Express Service
PM: Priority Mail
S2: Standard Mail Marketing
SA: Standard Mail Fulfillment

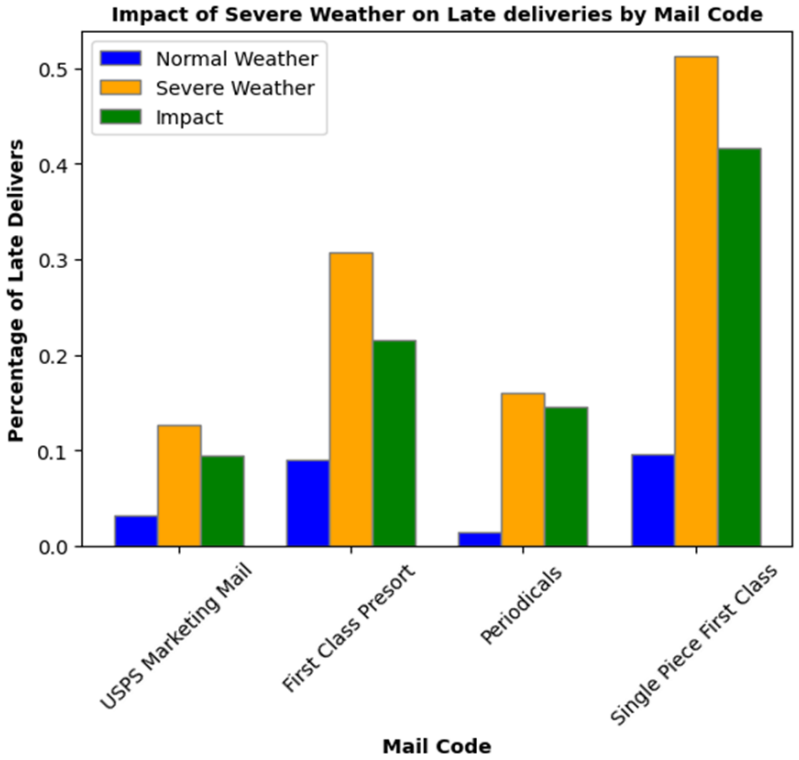
Exploratory Data Analysis- Mail



Exploratory Data Analysis- Mail



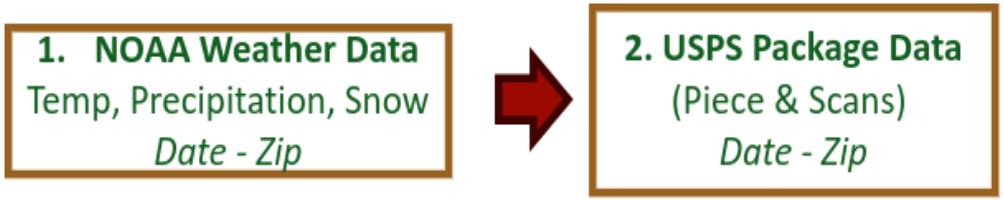
Exploratory Data Analysis- Mail



Model Selection

Data model: Packages

- **6k Unique Date and Zip Weather records, 4k Valid Weather API pulls from NOAA**
- **49k USPS Scans records, 37k USPS Piece records**
- **~17k, 15 cols Processed records for Model. Zip3 of '370', '371', '372' (Nashville Area) (Nashville Area)**



Scans (Feature Engineering)

time-delta-minutes: Time between the events 7, 10 (min and max). 7 Arrival at Unit, 10 Processed through USPS facility

Distinct_event_scans:
Counts of event scans (7, 10)

Processed Weather & USPS Data

	ServiceTypeCode	MailClassCode	Distinct_event_scans	time_delta_minutes	late	Zip_O	TMIN_O	TMAX_O	PRCP_O	SNOW_O	Zip_D	TMIN_D	TMAX_D	PRCP_D	SNOW_D
0	001	G0	1	0	0	37205	NaN	NaN	0.02	NaN	37205	NaN	NaN	0.02	NaN
1	001	G0	1	0	0	37221	5.0	28.0	0.00	NaN	37221	5.0	28.0	0.00	NaN
2	70	FC	1	0	0	37066	NaN	NaN	0.00	0.0	37066	NaN	NaN	0.00	0.0
3	001	G0	1	0	0	37087	13.0	27.0	1.09	4.0	37087	13.0	27.0	1.09	4.0
4	055	PM	1	0	0	37067	NaN	NaN	0.00	0.0	37027	NaN	NaN	0.00	0.0

Model Exploration Overview

Models	Accuracy	AUC	Precision	Recall	F1 Score
Random Forest Classifier	0.8903	0.9221	0.7914	0.6098	0.6888
XGBoost Classifier	0.8940	0.9327	0.8900	0.8900	0.8900
Gradient Boosting Classifier	0.8715	0.8977	0.8580	0.4247	0.5681
LightGBM Classifier	0.8886	0.9260	0.8315	0.5524	0.6628
CatBoost Classifier	0.8906	0.9271	0.8384	0.5581	0.6701
Neural Network	0.8629	0.8628	0.6887	0.5681	0.6226
Naïve Bayes	0.2482	0.5263	0.2076	0.9857	0.3430

XG-Boost Result Analysis: Packages

Hyperparameters:
Number of Trees: 100, 200, 300
Depth: 3, 4, 5
Learning rate: 0.1, 0.01, 0.2

Model Tuning:
Uses 5-fold cross-validation.
Parameter Grid Defined with GridSearchCV to find the combination of hyperparameters.

Best Model Parameters:
colsample_bytree: 0.9
learning_rate: 0.2
max_depth: 5
n_estimators: 300
subsample: 0.9

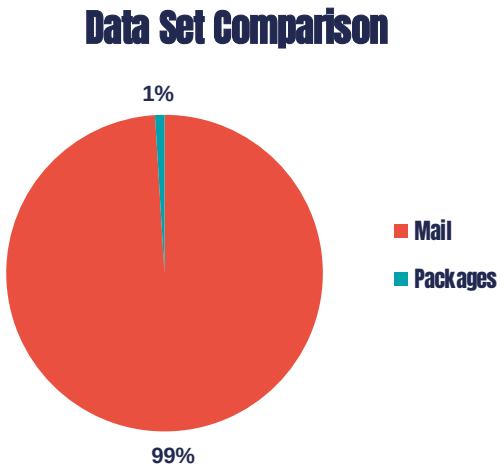
Best Cross-Validation Accuracy: 0.897

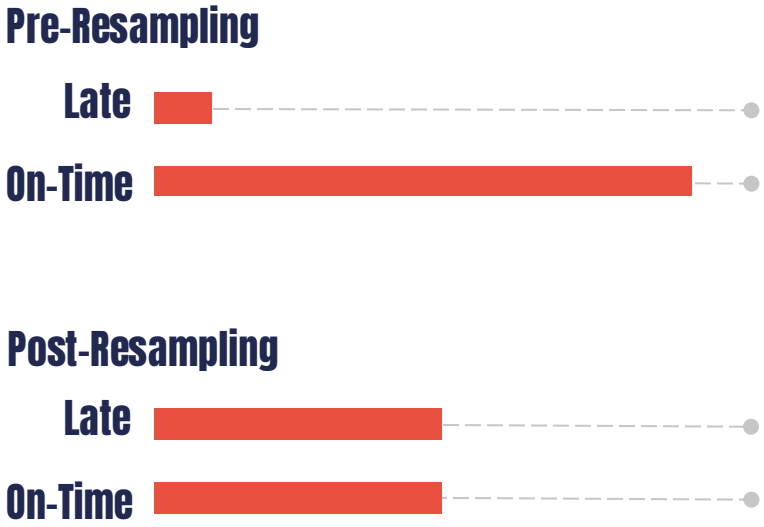
Model Performance:
Accuracy: 0.894, AUC: 0.933

TOP 5 Feature Importance
PRCP_D 0.13,
MailClassCode 0.13,
time_delta_minutes 0.12,
ServiceTypeCode 0.09,
TMIN_O 0.087

Model Validation

Preprocessing - Mail Data





Model Training - Mail Data

```
param_grid = {
'n_estimators': [100, 200, 300],
'learning_rate': [0.01, 0.1, 0.2],
'max_depth': [3, 4, 5],
'subsample': [0.7, 0.8, 0.9],
'colsample_bytree': [0.7, 0.8, 0.9]
}
```

Parameters	Package Model	Mail Model
colsample_bytree	0.9	0.9
learning_rate	0.2	0.2
max_depth	5	5
n_estimator	300	300
subsample	0.9	0.7
accuracy	0.897	0.937

Model Validation - Mail Data

Classification Report:

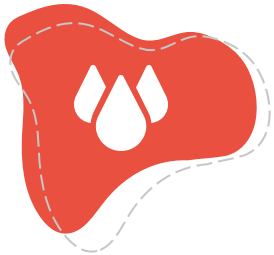
	Precision	Recall	F1-Score	Support
0	0.95	0.92	0.93	5017
1	0.92	0.95	0.93	4983
	Accuracy:		93%	
	Area Under the Curve:		98.6%	

Feature Importance:

Feature	Importance
d_SNOW	0.21
o_TMAX	0.19
d_TMIN	0.15
d_TMAX	0.12
o_PRCP	0.11

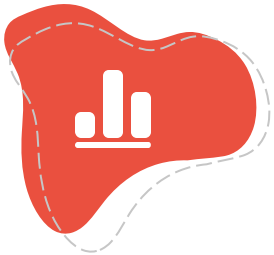
Project Summary

Key Take-Aways



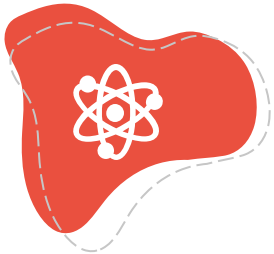
Weather

Severe weather **does** impact the delivery of mail & packages



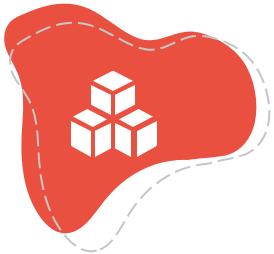
Mail Classes

Some classes of mail were more weather resistant than others



Model Performance

The XGBoost Model performed well on Mail & Package data



Data Quality

Clean and quality data is essential to model accuracy

Next Steps

Step 1

Attempt XGBoost Model on mail and packages with different originating and destinating locations

Step 2

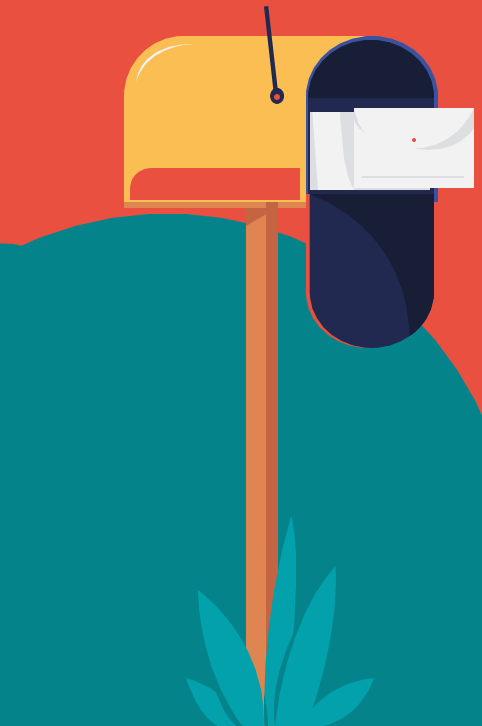
Conduct a pilot using a real-time severe weather event

Step 3

Enterprise wide implementation of solution and continued analysis

Thank You!

Questions?



Github

CREDITS: This presentation was created using a template by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**.