Remote Sensing for Forest Recovery

Team:

- Benjamin Frizzell
- Mavis Wong
- Hui Tang
- Zanan Pech

Mentor: Alexi Rodríguez-Arelis

Partner: Piotr Tompalski

Agenda

01 Problem Context & Data Source

02 Product Overview, Data Summary & Key Challenges

03 Data Science Techniques

04 Limitations and Potential Improvements

Problem Context & Data Sources

Background & Problem Context

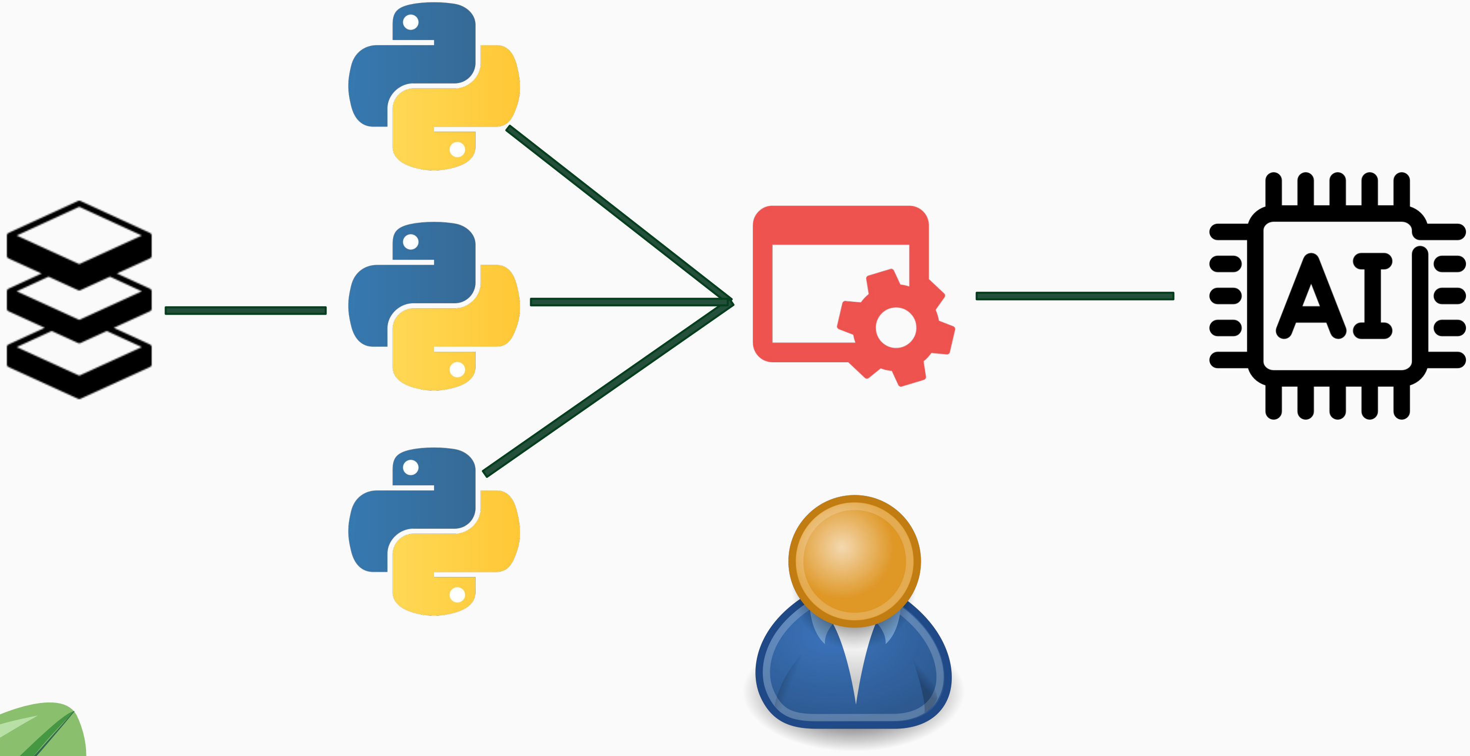
- Large scale afforestation efforts underway
- Newly planted trees are difficult to detect
- Need for scalable monitoring methods
- Data collected in Ontario by **Forest Canada**

Goals

- Discover whether remote sensing data are useful for predictions
- Identify the effective modelling approach
- The level of accuracy that can be achieved

Product Overview, Data Summary & Key Challenges

Data Product



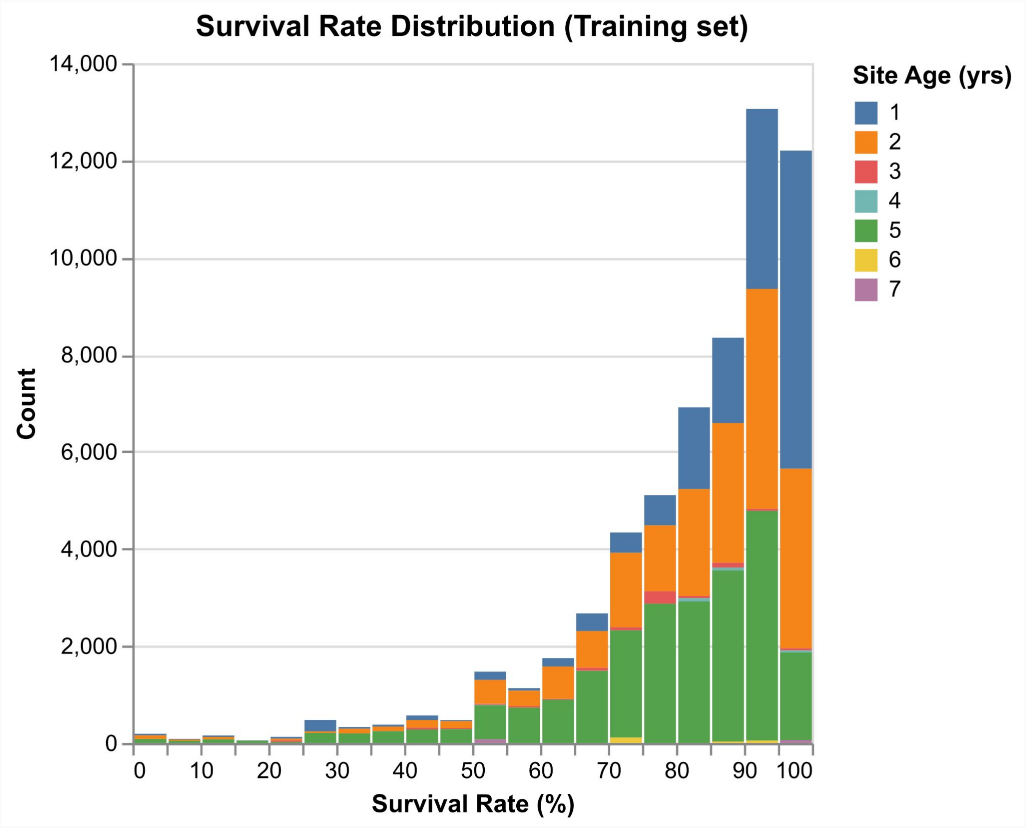
Data Overview

- Static Site Features
  - Area (Hectares)
  - Species Compositions
  - Species Types
  - Number of planted trees
  - Time of planting, survival assessment
- Target: survival rate (0 - 100%) measured across 7 years
- Spectral Indices: NDVI, NDWI , NBR etc.
  - Measures of greenness, soil exposure, water content, etc.
  - Measured approximately monthly

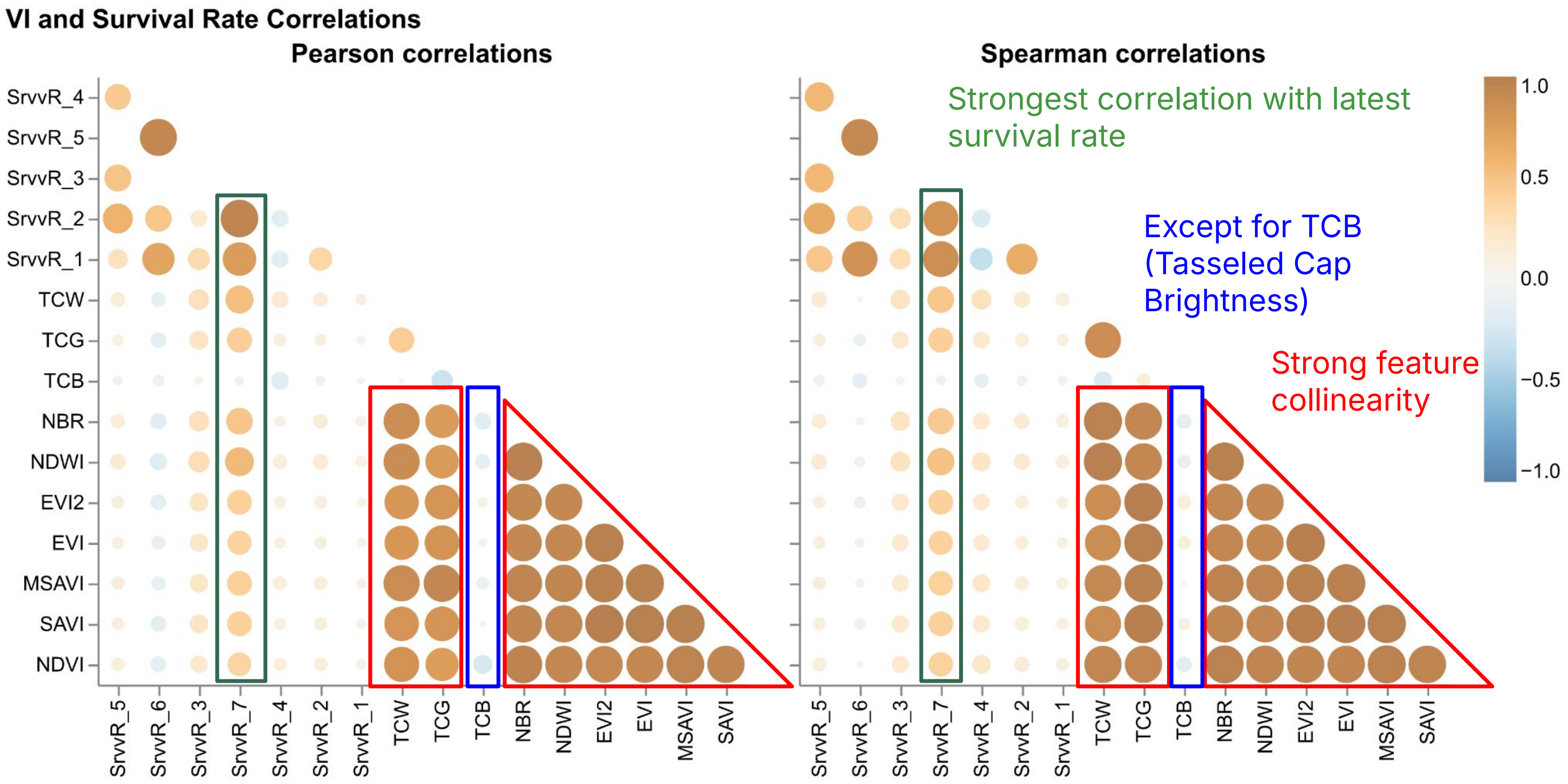
Age	Species Types	Assessment Date	...	Target
1	Conifer	2022-10-15	...	80
2	Deciduous	2023-10-15	...	60

Image Date	NDVI	NDWI	NBR	EVI	...
2018-10-15	1.0	0.8	0.6	0.9	....
2028-11-15	0.3	0.2	0.4	0.4	....

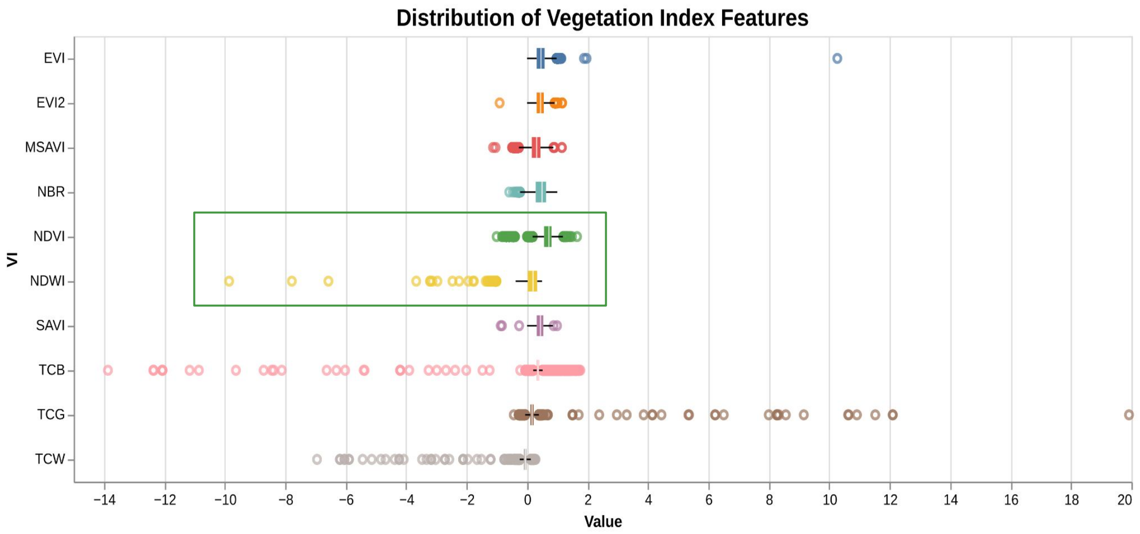
Challenges – Skewness

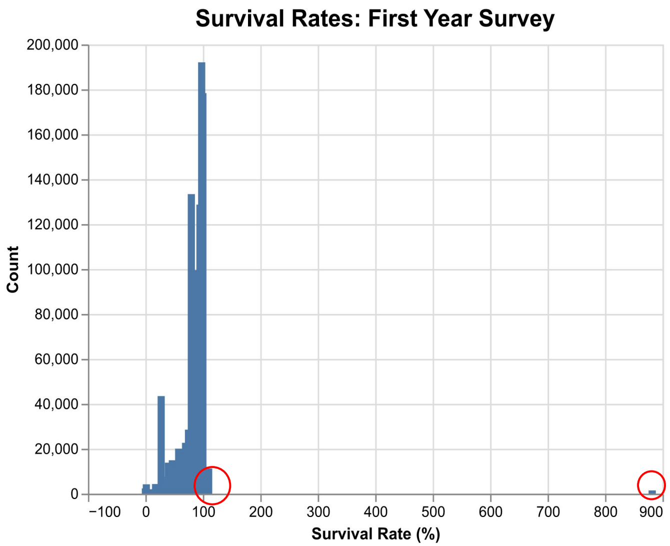


Challenges – High Correlation



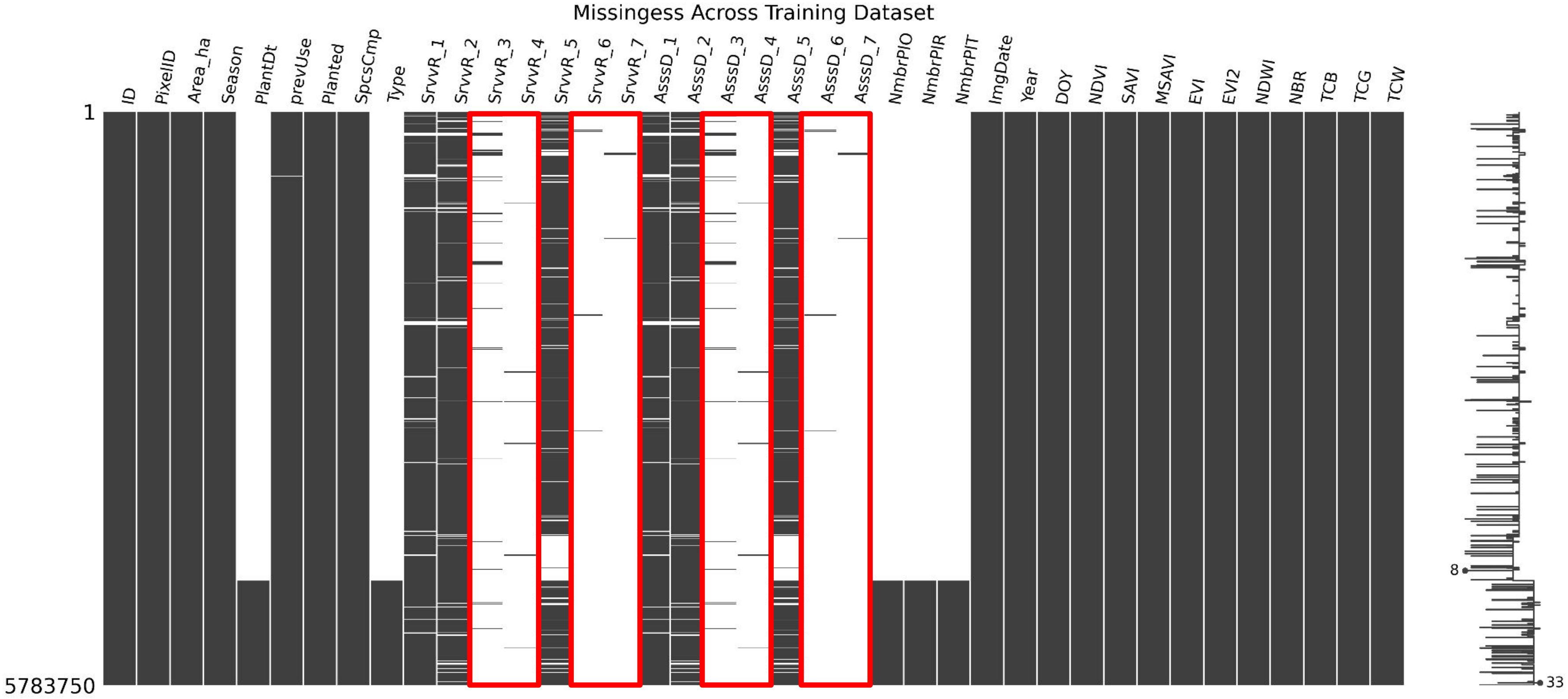
Challenges – Outliers





- Normalized indices (NDVI, NDWI) should not exceed [-1,1], outliers should be removed
- Some survival rates in first survey exceed 100%, should be removed

Challenges – Missing Values



Missing target survival rate/record time

Challenges – Thresholds (Binary Conversion)

50%

60%

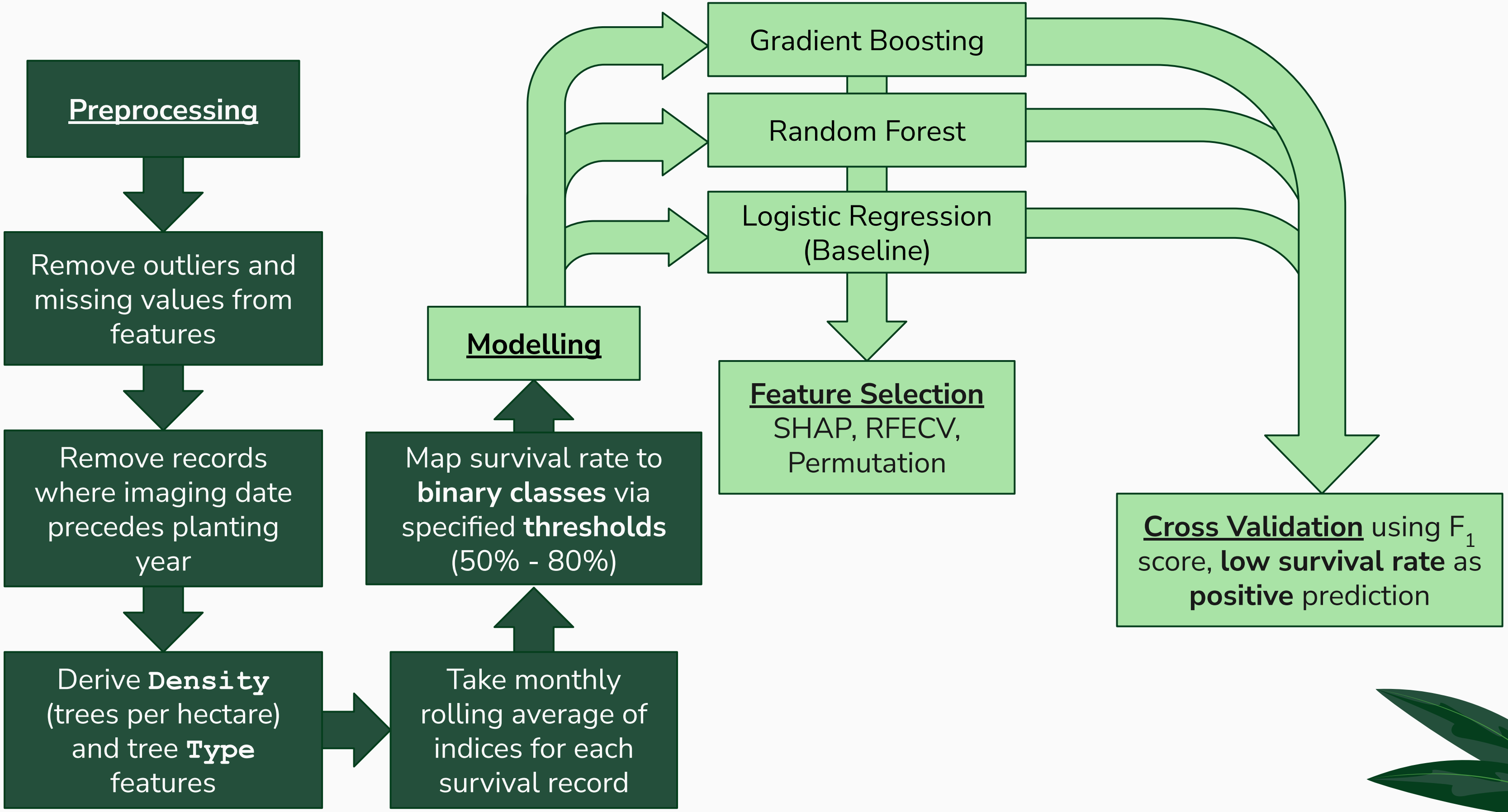
70%

80%

Age	Assessment Date	Survival Rate	...	Target
1	2022-10-15	60	...	HIGH
2	2023-10-15	40	...	LOW

Data Science Techniques

Phase 1 : Classical Modelling Pipeline

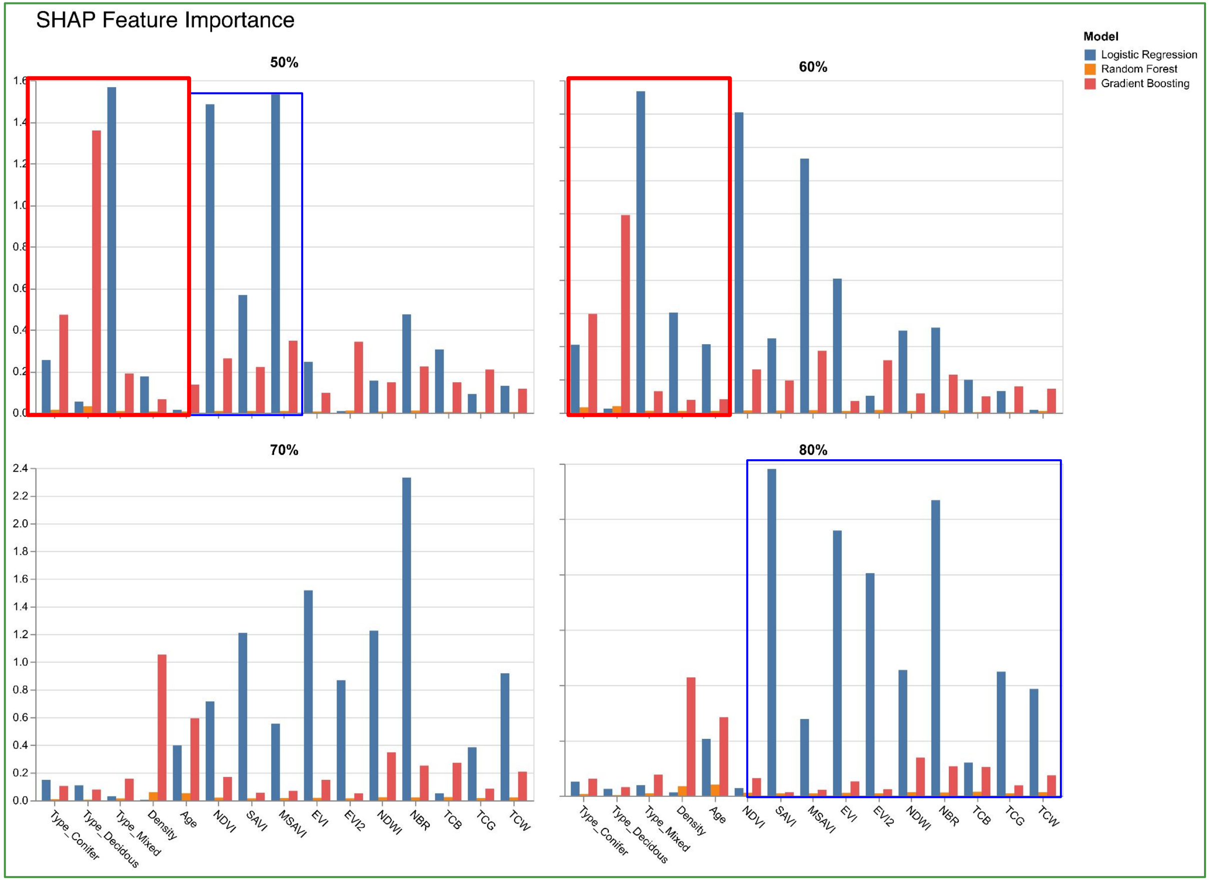


Permutation Importance

- Randomly shuffle a feature and see how much performance decreases
- fewer features affect performance as threshold decreases
- Density and age show high importance generally for tree models
- Importance depends on both model and threshold

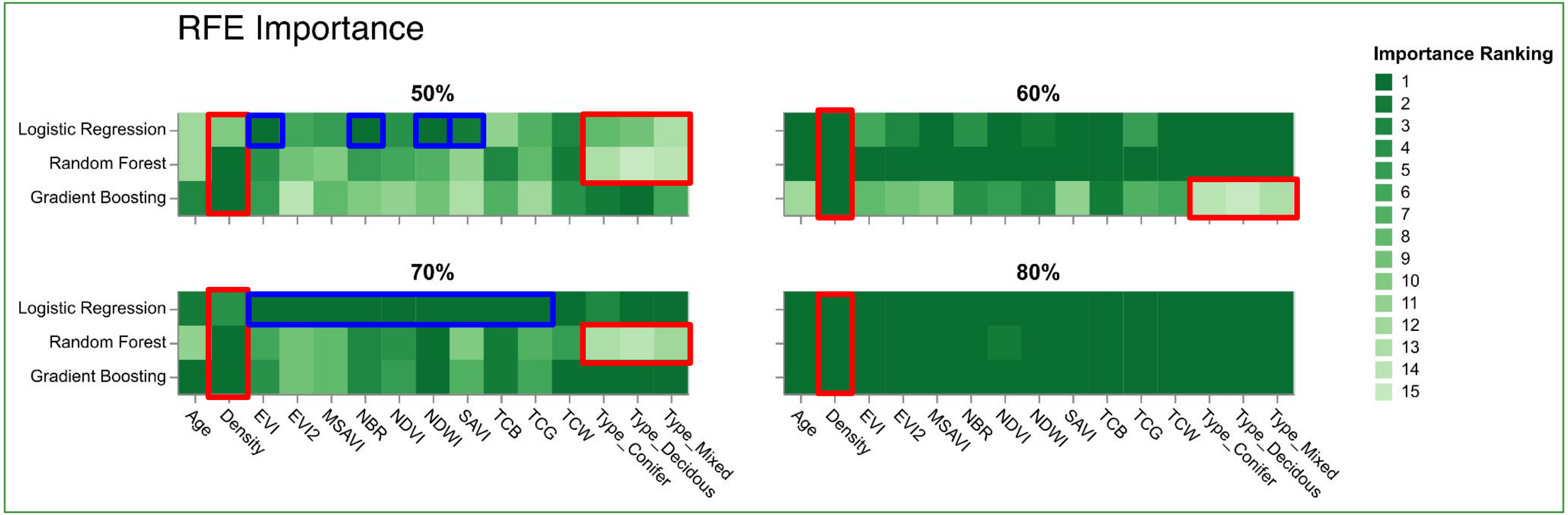
SHAP Importance

- Train model on all feature subsets, compute weighted average change to log-odds from adding that feature
- Higher contribution from more remote sensing features as threshold increases
- Type features have strong contribution at low thresholds, Density and Age have lower contribution (contrary to permutation)
- No strong contribution measured by Random Forest



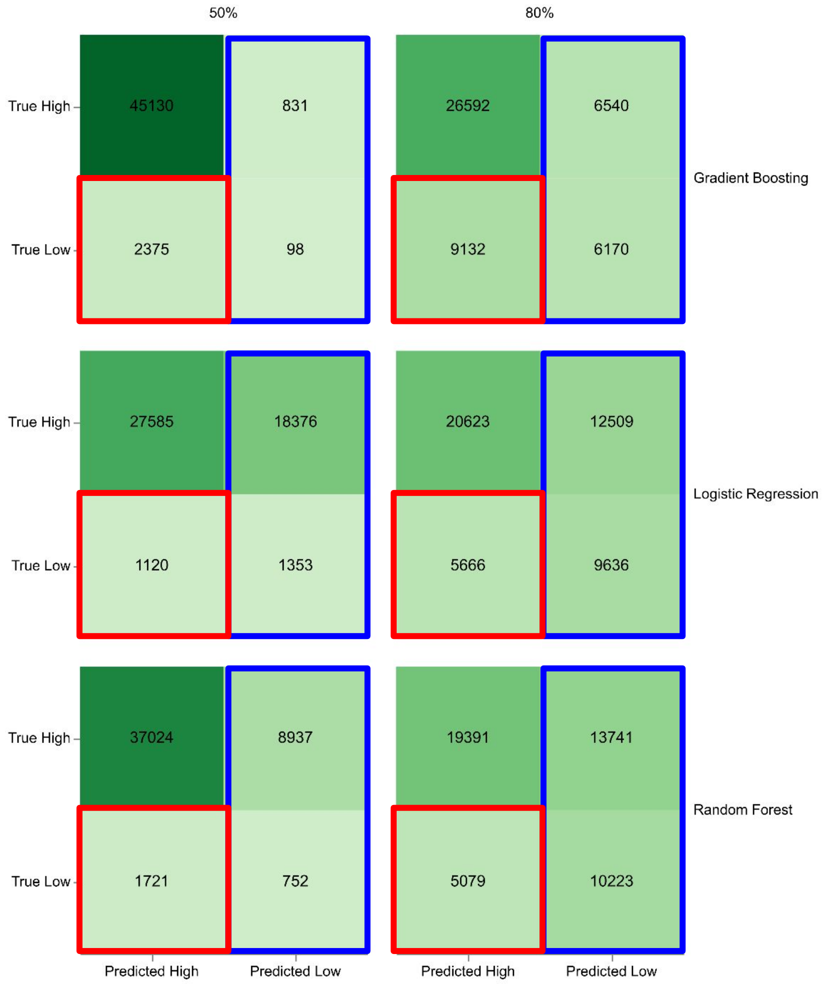
RFE Rankings

- **Type shows little use**, especially at low thresholds
- **Density consistently helpful** across most thresholds
- Less features may be be needed for identifying low survival rates (EVI, NBR, NDWI, SAVI)
- Importance varies greatly with **threshold** and **model**, continue with **all features** to be safe



Confusion Matrices

- Increasing threshold leads to more balanced data, more true positive predictions, relatively fewer false positives
- But false negatives also increase rapidly, suggesting issues **beyond class imbalance**



Phase 1: Conclusion

- Random Forest and Gradient Boosting **cannot significantly outperform** baseline

- Class imbalance and data processing likely limiting model performance: losing vital information by averaging over time

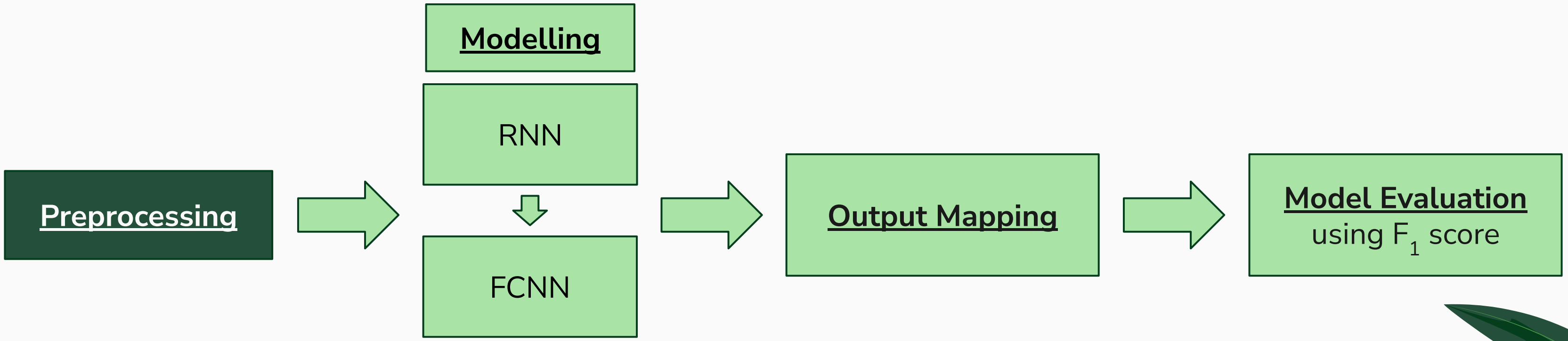
- Models that are more suitable for **time series** may be necessary!

	50% Threshold			80% Threshold		
Model	F_1 score	Precision	Recall	F_1 score	Precision	Recall
Gradient Boosting	0.058	0.105	0.040	0.411	0.485	0.403
Random Forest	0.124	0.078	0.304	0.521	0.427	0.668
Logistic Regression	0.122	0.069	0.547	0.515	0.435	0.630

Phase 2 : RNN Modelling Pipeline

Recurrent Neural Network (RNN) :

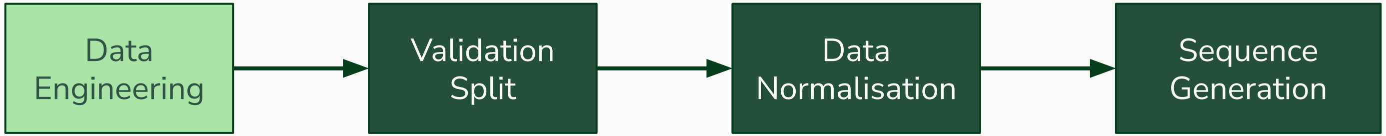
- Train RNN Regression Model → Map output to binary class for evaluation
- No defined classification threshold → Avoid training multiple RNN models



Data Engineering

- **Log Transformed time_delta**
  - Time difference between image date and survey date
  - Capture irregularity in time intervals of the satellite records

- **Negative cosine transform DOY**
  - $-\cos(2\pi \times DOY/365)$
  - Capture seasonality in satellite signals



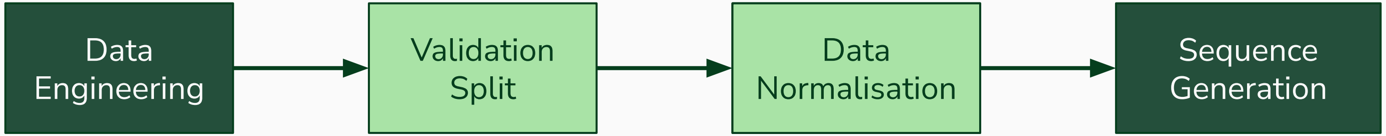
Data Preprocessing

Validation Split

- Split the test data to get validation set
- Use during training to validate model performance

Data Normalization

- RNN sensitive to data scaling
- Avoid vanishing/exploding gradient



Sequence Generation

Survival Records

ID	PixelID	SrvvR_Date	Age	target
1	11	21/05/21	1	80
2	12	10/11/13	2	70
5	55	01/01/22	6	80

Imaging Records

ID	PixelID	ImgDate	DOY	NDVI	…	TCW
3	13	01/03/14	60	0.6	…	-0.3
4	14	07/04/21	97	8.6	…	0.5
…	…	…	…	…	…	…

Sequential Data fed to RNN

ID	PixelID	ImgDate	time_delta	DOY	NDVI	…	TCW
5	55	01/01/18	1461	1	0.7		0.6
5	55	01/02/18	1430	32	0.2		0.8
5	55	01/07/19	915	182	-0.3		-0.7

FOR each row in Survival Rates table:

- Search image table for records with matching **`(ID, pixelID)`**
- Select all records up till survey date **`(ImgDate ≤ SrvvR_Date)`**

RNN Modelling

GRU
(Gated Recurrent Unit)

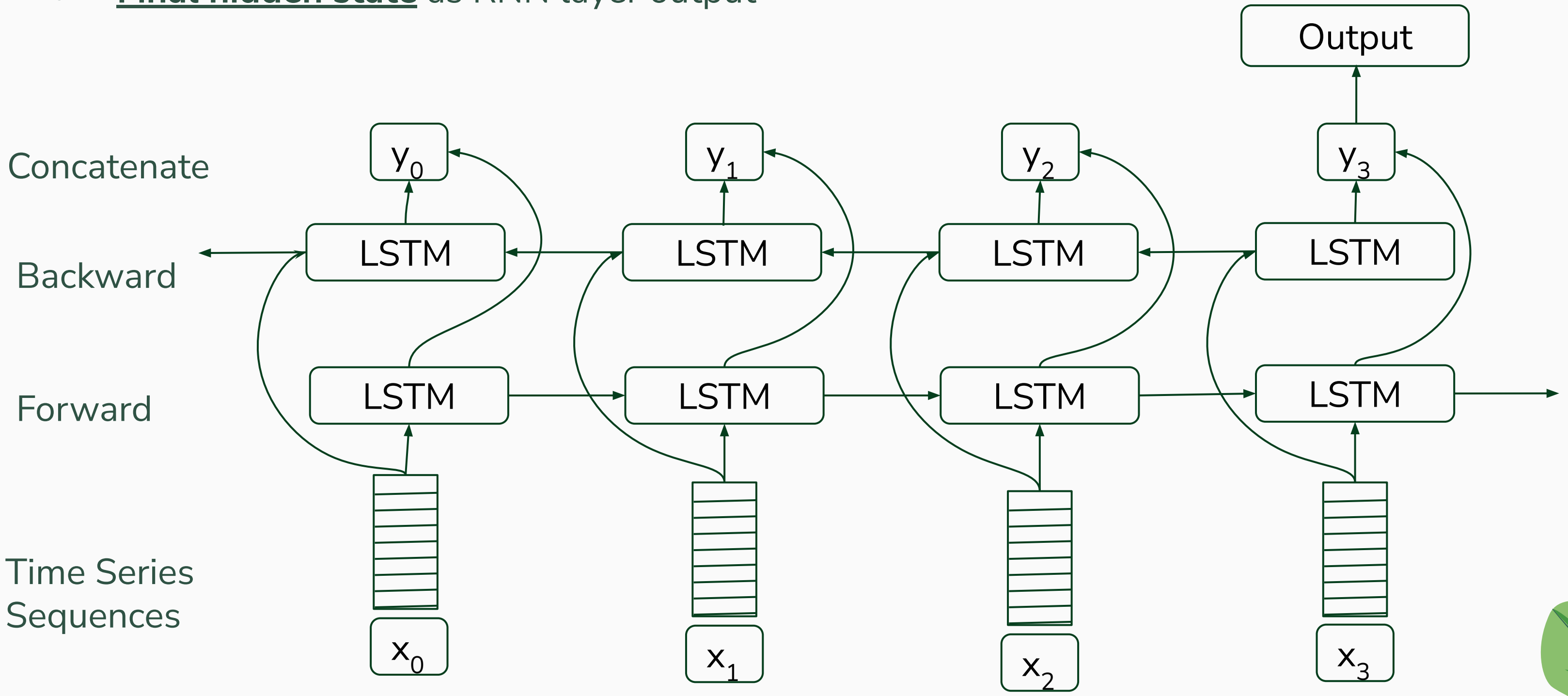
- Simplified LSTM
- Good at capturing short- and mid-term dependencies.
- Faster training

LSTM
(Long Short–Term Memory)

- More complex architecture
- Good for capturing long term dependencies
- More robust for complex or longer sequences.

RNN Architecture

- **Bi-directional RNN**: capture long-term time dependencies more effectively
- **Final hidden state** as RNN layer output

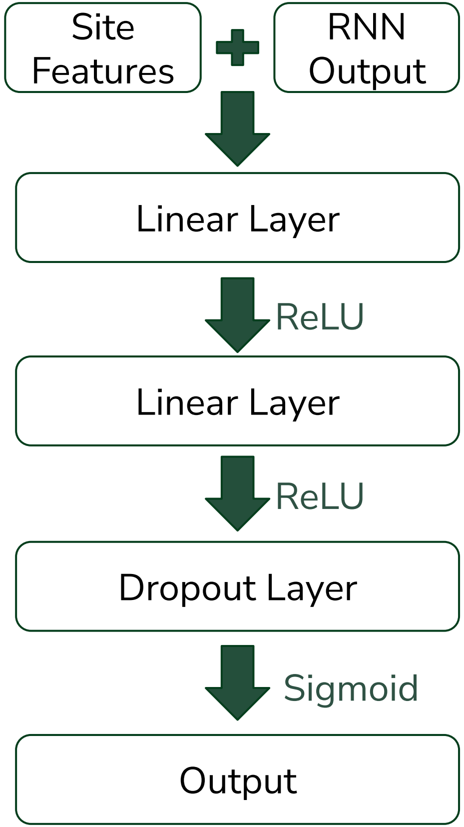


Feed–Forward Layer

1. **Concatenate :** RNN output (+ site features)
2. **Linear layer + ReLU activation** : Nonlinearity
3. **Dropout Layer :** Avoid Overfitting
4. **Sigmoid activation**: map output to desired range

<u>**Loss Function**</u>

- **MSELoss** : penalises large error
- Data skewed towards high survival rates
- Improve predictions for low survival rate records



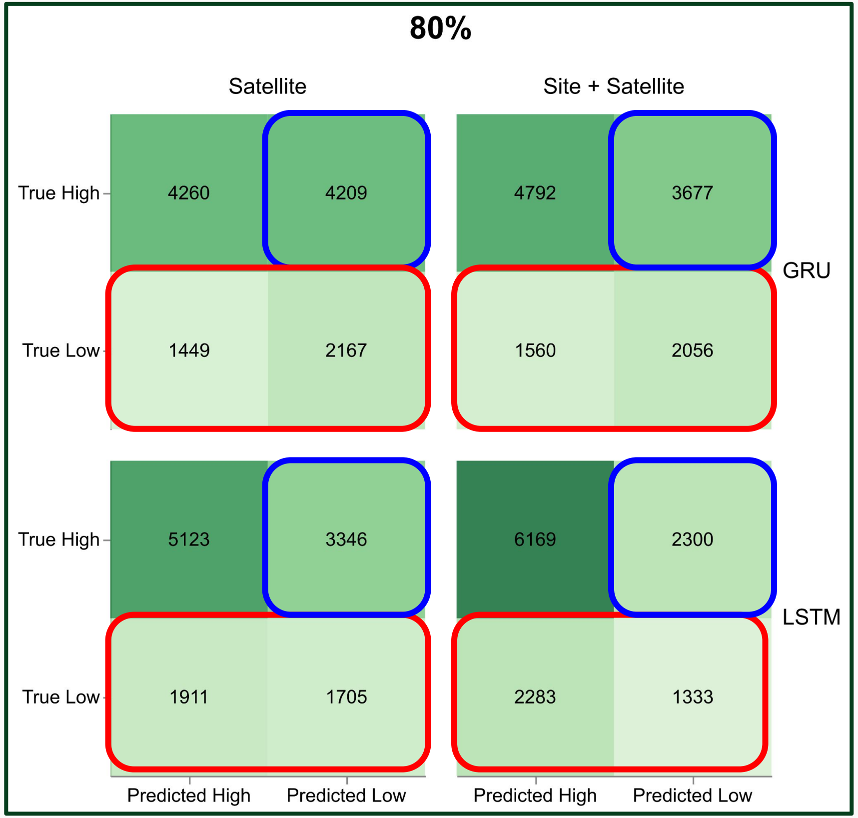
RNN Model Evaluation

		50%			80%		
Model	Features	F_1 Score	Precision	Recall	F_1 Score	Precision	Recall
LSTM	Site + Satellite	0	0	0	0.368	0.367	0.369
	Satellite	0	0	0	0.393	0.338	0.472
GRU	Satellite	0	0	0	0.434	0.34	0.599
	Site + Satellite	0	0	0	0.44	0.359	0.569

- Fails to make any correct positive prediction

- Much better performance
- Fails to outperform classical models

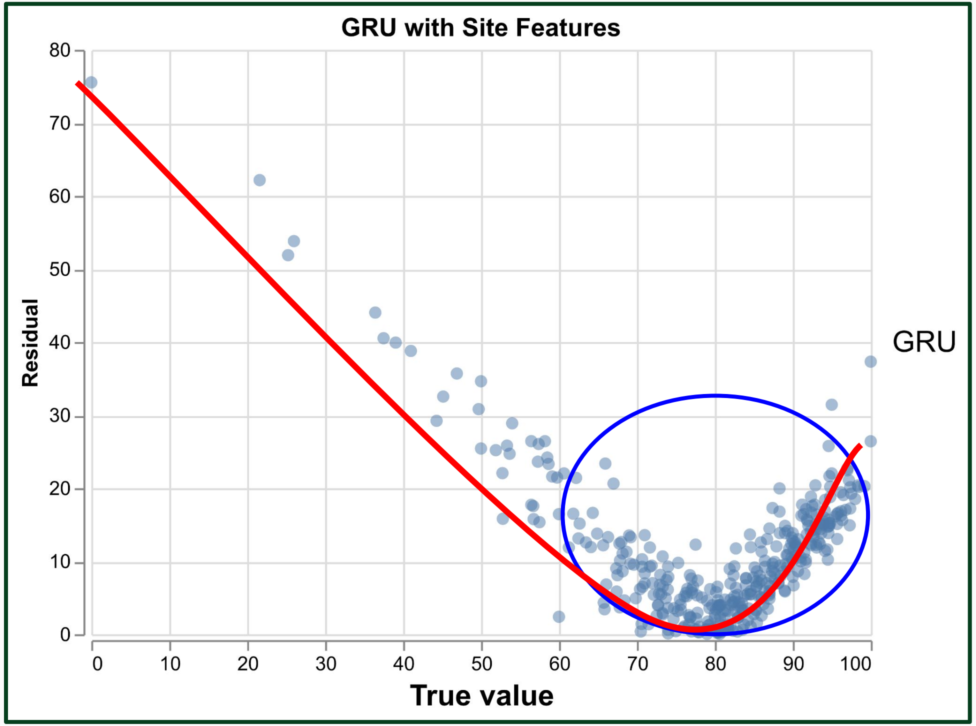
Confusion Matrices



- GRU performs better than LSTM
- Higher True positive
- Lower False Negative

- But this also comes with high False Positive ...

Residual Plots



- Imbalance data, where most true value is >70%

- Relationship resembles a convex function centered at 80%
- Model is not making useful predictions, predicting a survival rate ~80% most of the time

Limitations and Potential Improvements

Recap: Current Limitations

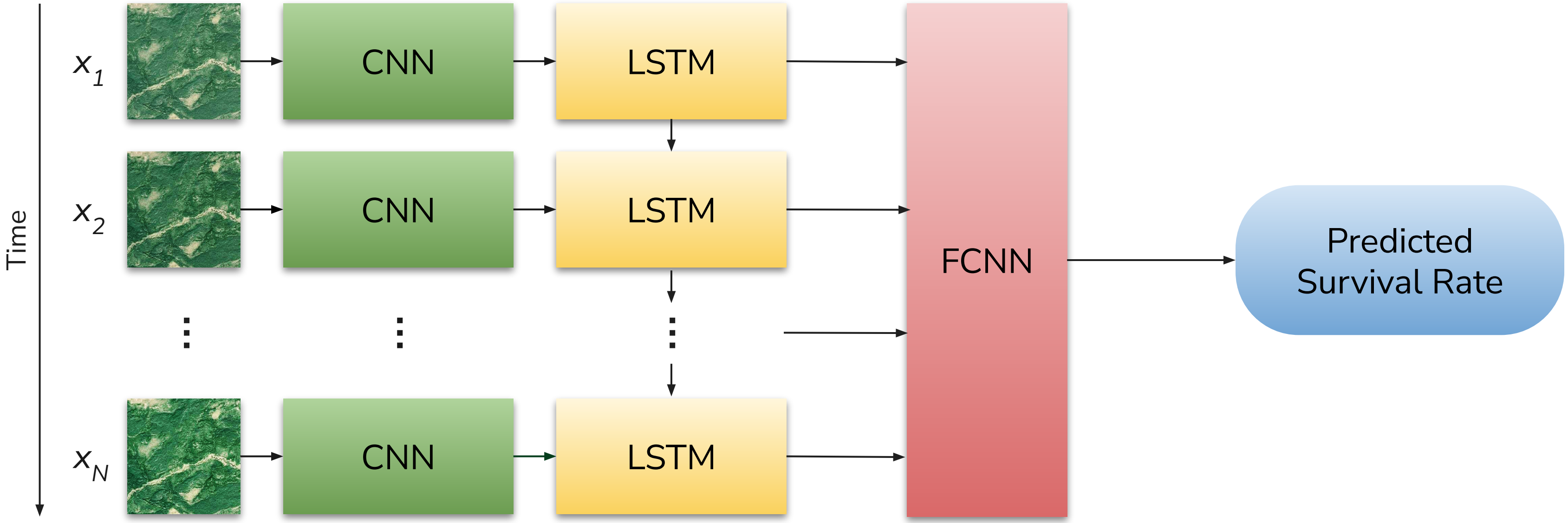
- **Imbalance:**
  - Lack of data on sites with low survival rates leads to biased predictions
- **Classical Modelling:**
  - Fails to capture **sequential changes** in vegetation indices **across time**
- **RNN Modelling:**
  - Can handle temporal structure
  - Fails to capture **spatial correlations** between **pixels within sites**
- **Misleading Target Labels:**
  - Survival rates are recorded per site, but predictions made per pixel
  - ‘Healthy’ and ‘unhealthy’ pixels in same site will have same target

Recommendations – Data

- **Use higher resolution satellite data**
  - Capture finer details, direct imaging should work with similar models!
  - Sentinel-2 offers 10m x 10m resolution data
  - Planet: 3m x 3m resolution, 15 day imaging cycle
  - Drone imaging: cm level resolution
- **Obtain more fine-grained survival records**
  - Per-pixel rather than per site for more consistent targets
- **Incorporate spatial data, make site-level predictions**
  - e.g. GPS coordinates, geodata etc.
  - Learn spatial and temporal patterns, relationships between pixels.
- **Obtain consistent annual survival records**
  - Improve model performance with more consistent temporal data

Potential Next Steps: CNN-LSTM

- **Convolutional Neural Network (CNN):** Used for image processing; extract features and spatial relationships from vegetation indices within site polygon at each time step
- Pass each CNN output to **LSTM** or **GRU** for sequence processing, then to **fully connected neural network (FCNN)** for prediction



Thank You