AIAP Batch 15 Technical Assessment

Billy Lim Jun Ming

Exploratory Data Analysis (EDA)

EDA

Data Cleaning

- Merged pre-purchase & post-trip data
- Based on Ext_Intcode
  - Rows with duplicate Ext_Intcode were dropped
  - Kept the row with less NA values (or the later row)

index	Cruise Name	Ticket Type	Cruise Distance	Ext_Intcode	WiFi	Dining	Entertainment
89340	Blastoise	None	150 KM	BL100AELMIT	0.0	1	1.0
89343	Blastoise	Luxury	150 KM	BL100AELMIT	0.0	1	1.0
44849	Blastoise	Luxury	1464 KM	BL100AQXMUS	1.0	0	1.0
15647	Lapras	Standard	1733 KM	BL100BAEEDV	NaN	1	NaN
15642	Lapras	Standard	1733 KM	BL100BAEEDV	NaN	1	NaN

EDA

Data Cleaning

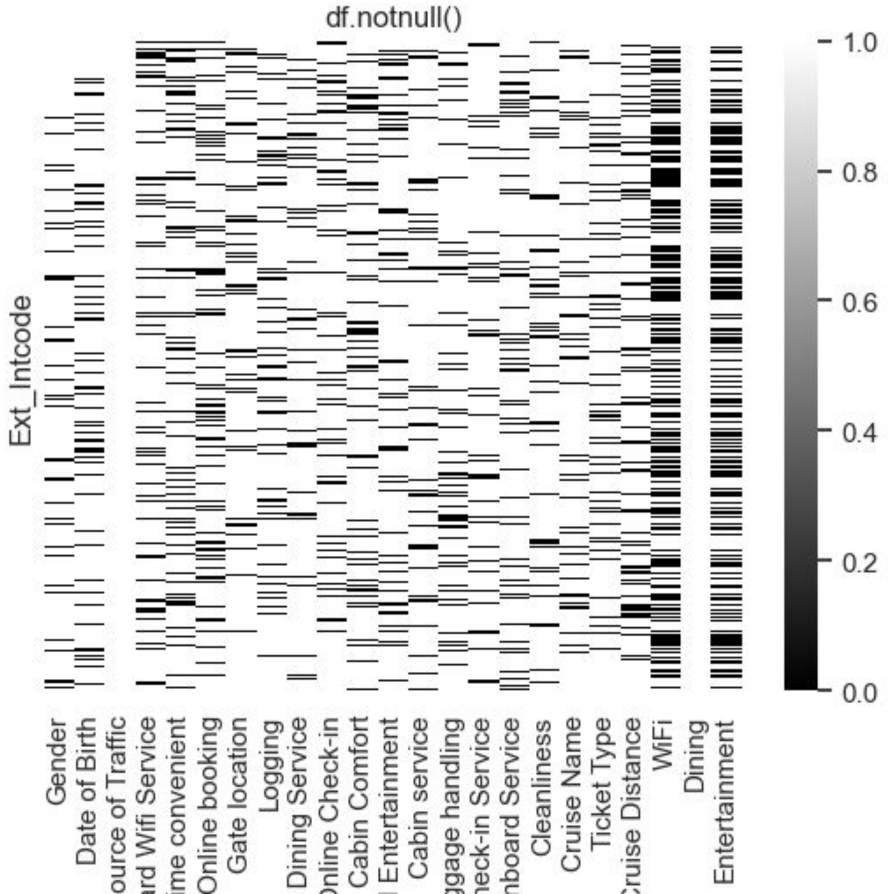
- Fixed typos in Cruise Name
  - To Blastoise or Lapras
  - Based on Levenshtein edit distance
- Typecasted appropriately
  - Ratings → Ordinal Numbers
  - Dates → Datetimes
  - km, miles → km

index	Cruise Name	Ticket Type	Cruise Distance	Ext_Intcode	WiFi	Dining	Entertainment
0	Blastoise	None	3567 KM	LB446RWOOZI	1	1	1
2	IAPRAS	Deluxe	1167 KM	BL713UHBAAN	NaN	0	0
3	Lapras	Deluxe	280 KM	LB243DMKCFL	NaN	0	1
9	None	Luxury	None	LB251DCACEW	0	0	1
12	blast	Standard	236 Miles	LB810DDUDEB	NaN	0	NaN
26	lap	Luxury	331 Miles	LB994CFCVQZ	0	0	1
37	blastoise	Standard	1085 KM	BL870JKZNZY	NaN	0	NaN
42	blast0ise	None	366 KM	BL332YRXJQW	NaN	1	NaN
45	lapras	Luxury	163 KM	LB265JZQPLM	0	1	0

EDA

Post-Trip Satisfaction Survey

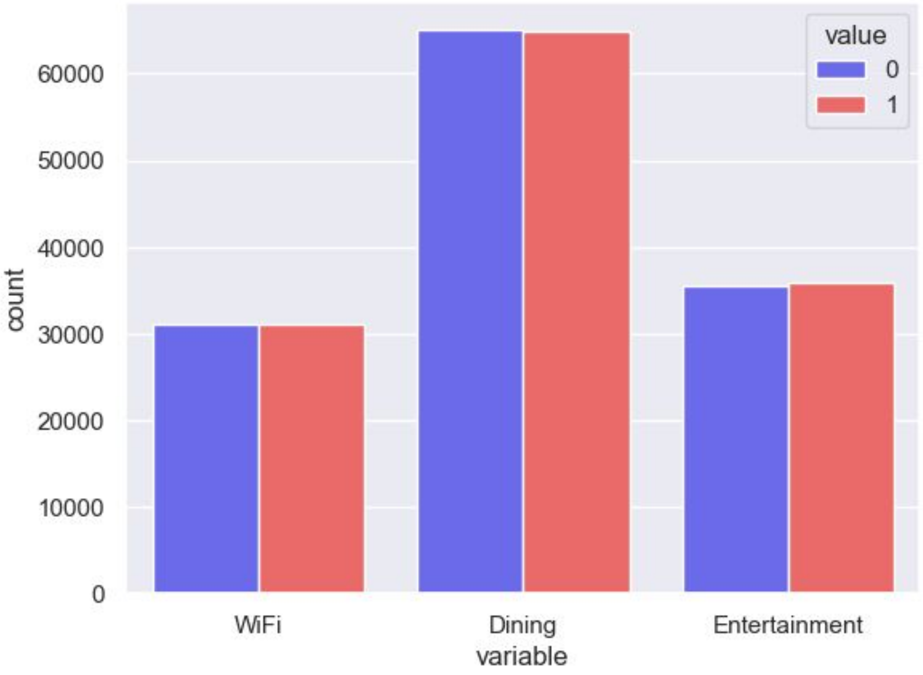
- NA values are dubiously distributed
  - Many NA for WiFi & Entertainment
  - Zero NA for Dining



EDA

Post-Trip Satisfaction Survey

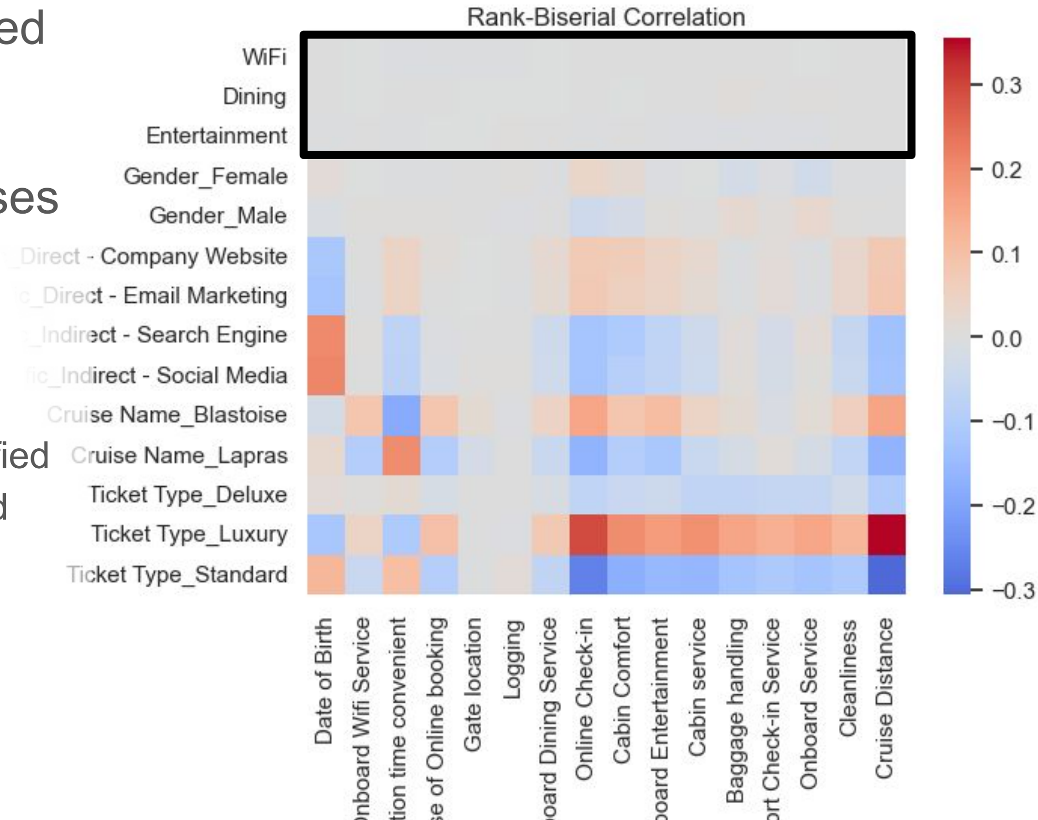
- NA values are dubiously distributed
  - Many NA for WiFi & Entertainment
  - Zero NA for Dining
- Satisfied and dissatisfied responses were perfectly balanced



EDA

Post-Trip Satisfaction Survey

- NA values are dubiously distributed
    - Many NA for WiFi & Entertainment
    - Zero NA for Dining
- Satisfied and dissatisfied responses were perfectly balanced
- Uncorrelated with pre-purchase importance ratings
    - WiFi rated unimportant → 50% satisfied
    - WiFi rated important → 50% satisfied
- **Dropped post-trip survey data**



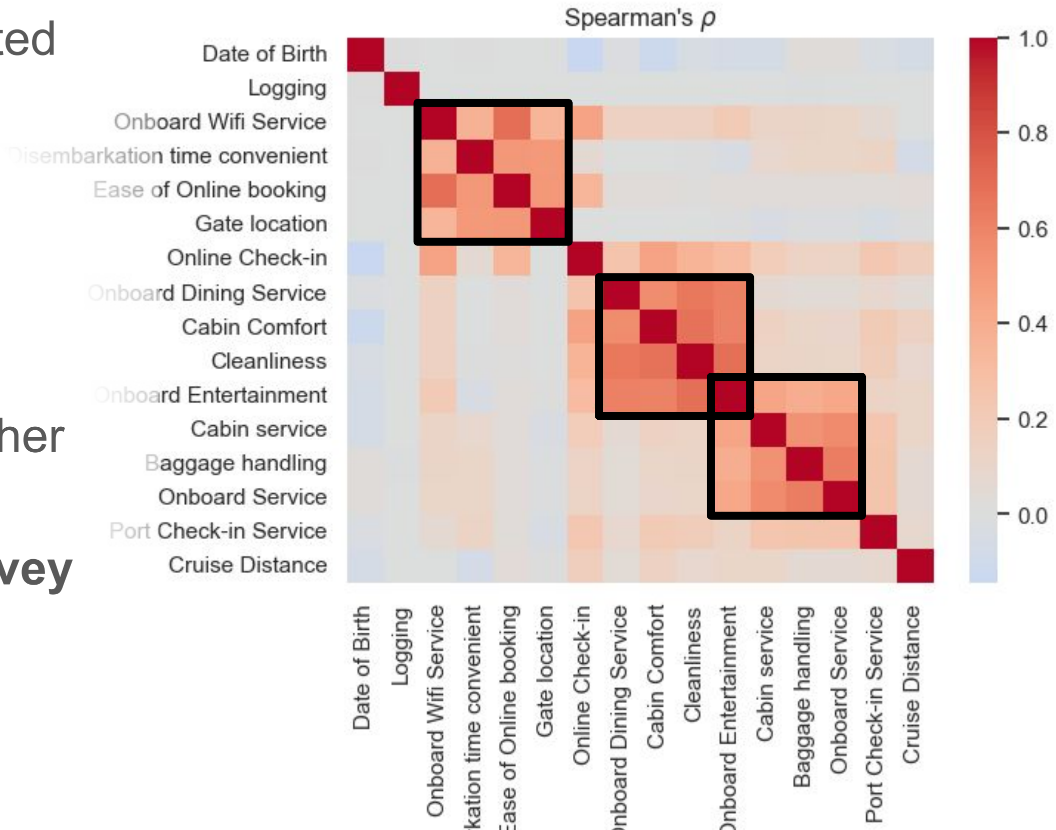
EDA

Pre-Trip Importance Survey

- Subsets of criteria are correlated
  - Convenience factors
    (WiFi, embarkation timing & gate)
  - Onboard facilities
    (cabin comfort, dining, cleanliness)
  - Hospitality services
    (Baggage handling, onboard service)

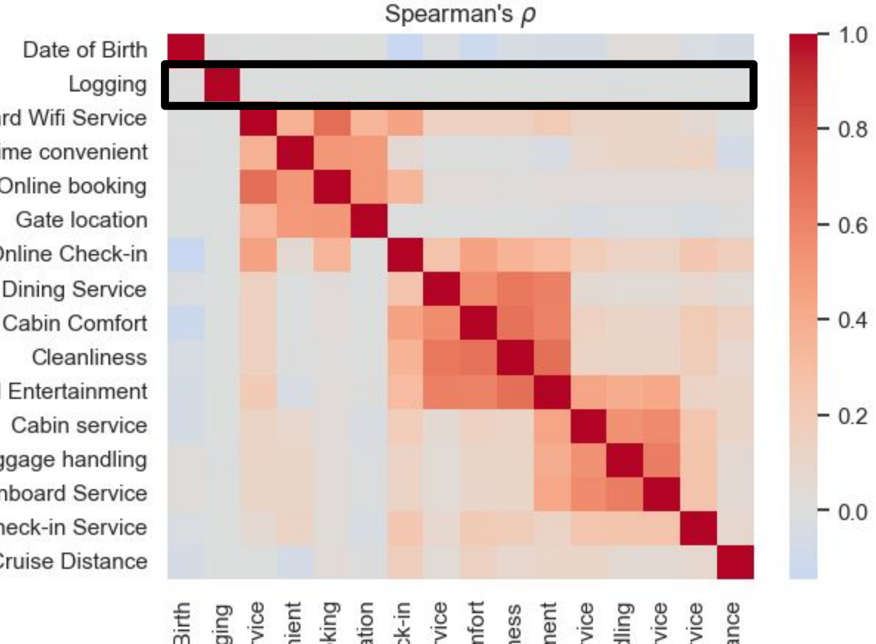
- We can aggregate them together

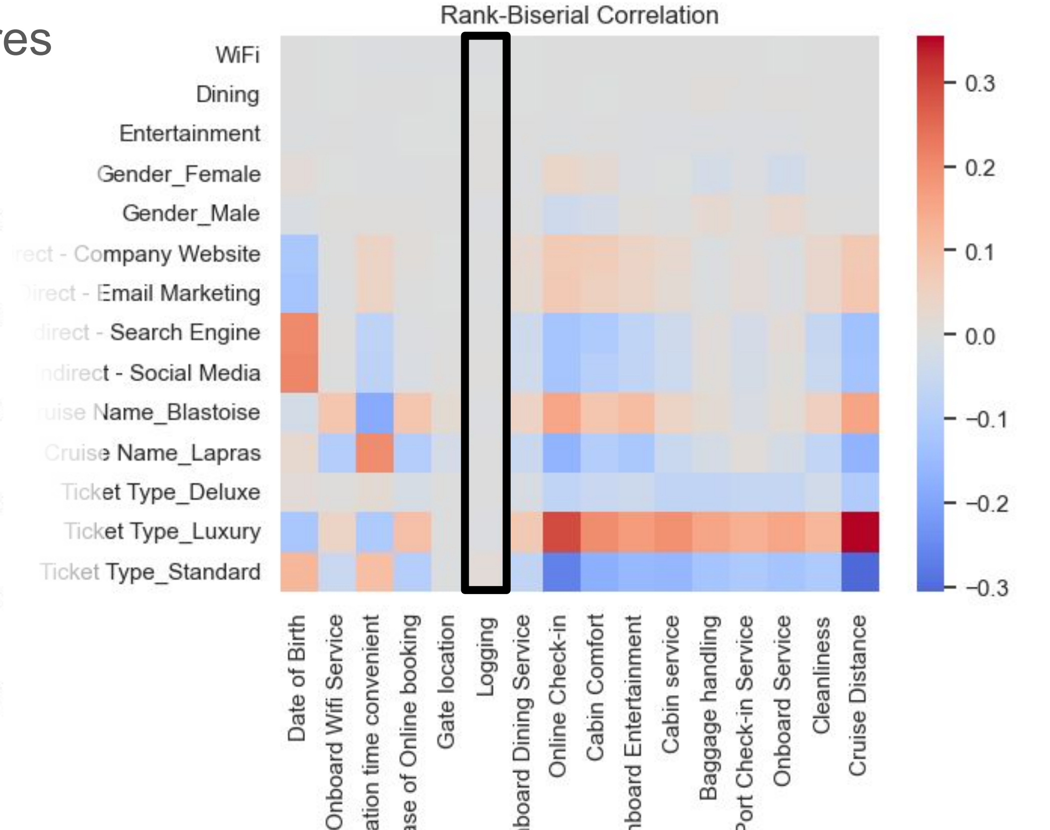
- **Applied PCA on pre-trip survey**



Logging

- Uncorrelated with all other features
- **Dropped logging data**

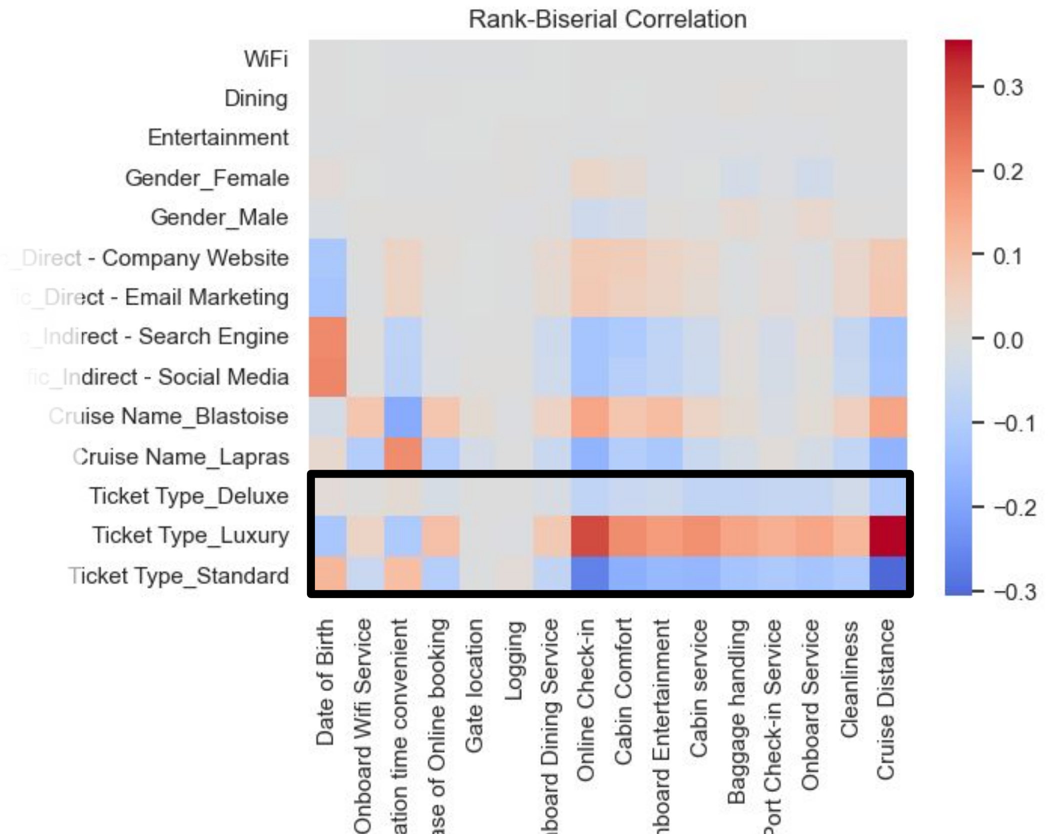




EDA

Ticket Type

- Luxury Tickets
  - Older age group
  - More exacting
- Standard Tickets
  - Younger age group
  - Less exacting
- Deluxe Tickets
  - Broader age group
  - Middle ground



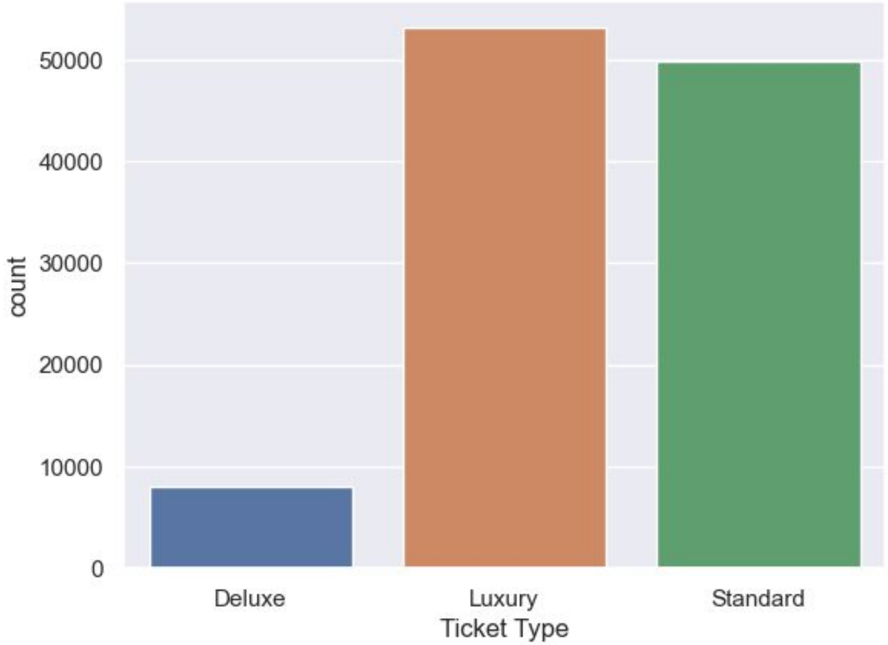
EDA

Ticket Type

- Very imbalanced distribution
  - Deluxe tickets make up <10% of sample

Since it's our **target variable**, we must

- Use stratified splittings
- Choose scoring metric appropriately



ML Models & Performance

ML Model

Model Pipeline

<table>
<tr><th>Ticket Type</th><th>Gender</th><th>Cruise Name</th><th>Source of Traffic</th><th>Date of Birth</th><th>Cruise Distance</th><th>Pre-Trip Importance Survey</th><th>Logging</th><th>Post-Trip Satisfaction Survey</th></tr>
<tr><td rowspan="3">Label Encoder</td><td colspan="3">-</td><td>Year()</td><td colspan="2">-</td><td colspan="2" rowspan="3">Dropped</td></tr>
<tr><td colspan="3">Impute Most Frequent Category</td><td colspan="3">Impute Mean</td></tr>
<tr><td colspan="3">One-Hot Encode</td><td colspan="2">-</td><td>PCA</td></tr>
<tr><td colspan="9">ML Model</td></tr>
</table>

Chosen ML models:

- Linear Support Vector Machine *(simple, explainable)*
- Random Forest *(ensemble, minimise variance)*
- Gradient-Boosted Tree *(ensemble, minimise bias)*

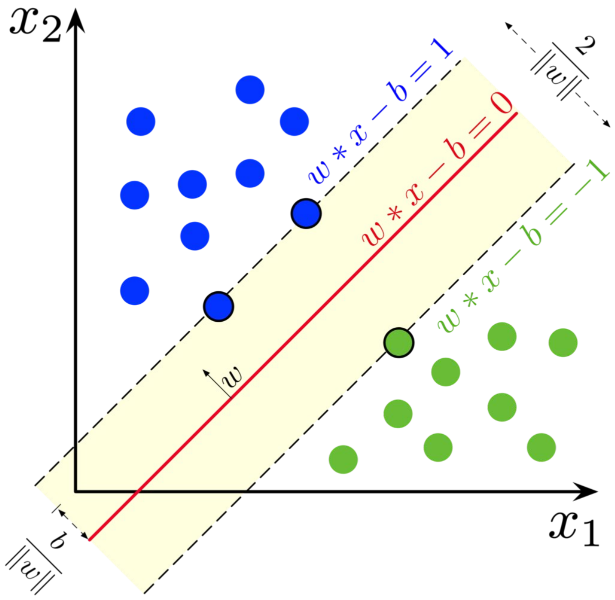
ML Model

ML Models

Chosen ML models:

- Linear Support Vector Machine

  (simple, explainable)

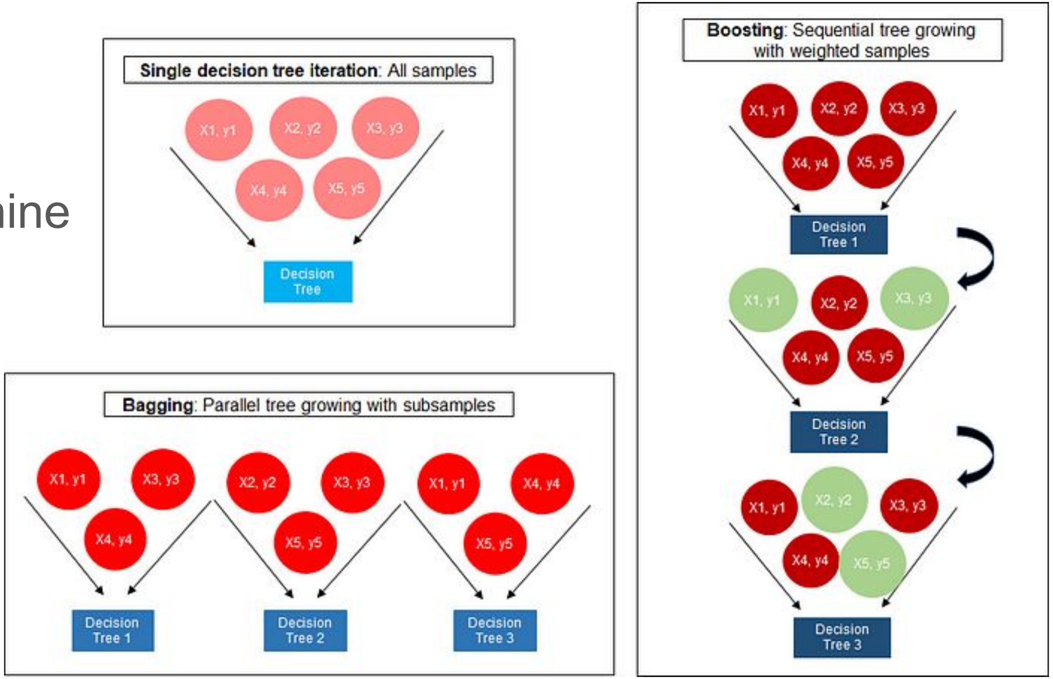


Source: https://en.wikipedia.org/wiki/Support_vector_machine

ML Models

Chosen ML models:

- Linear Support Vector Machine
  (simple, explainable)
- Random Forest
  (ensemble, minimise variance)
- Gradient-Boosted Tree
  (ensemble, minimise bias)

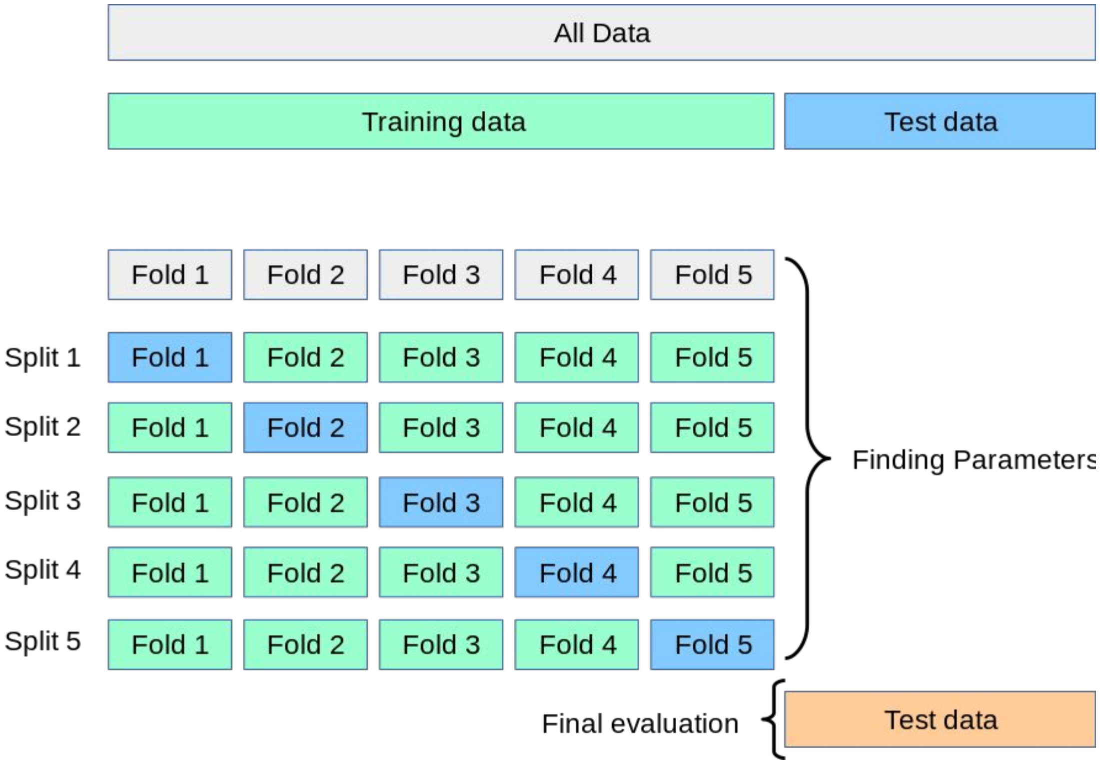


Source: https://towardsdatascience.com/the-ultimate-guide-to-adaboost-random-forests-and-xgboost-7f9327061c4f

ML Model

Training

- Stratified train-test split
- Stratified 5-fold cross-validation



Source: https://scikit-learn.org/stable/modules/cross_validation.html

ML Model

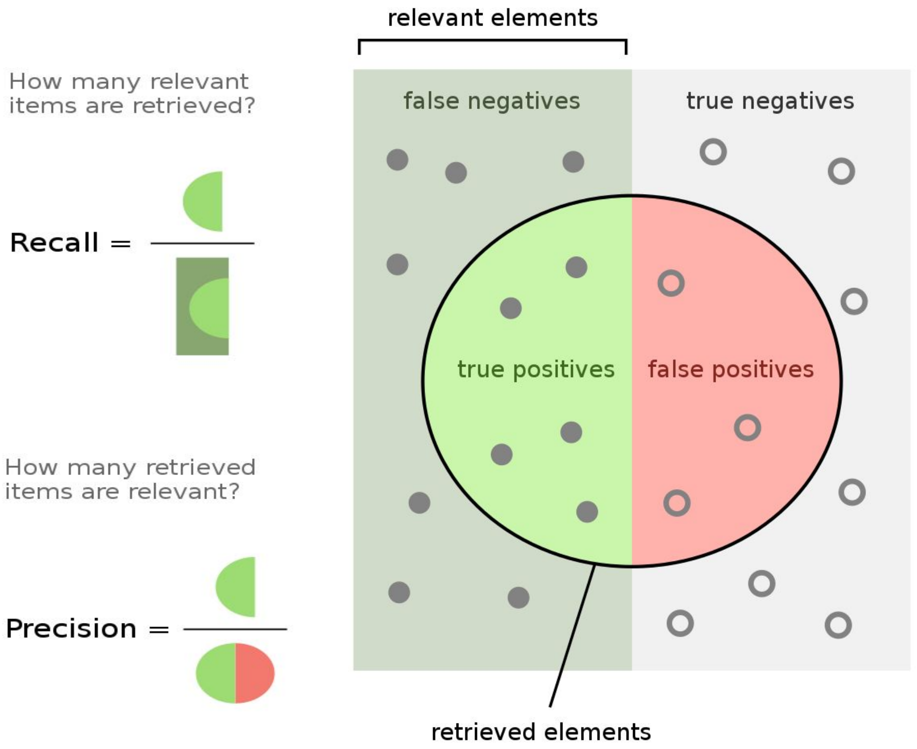
Scoring

- F1 score

$$F_1 = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

- Macro-averaged

$$F_1\text{-macro} = \frac{1}{3}\begin{bmatrix} F_1(\text{Luxury}) \\ +F_1(\text{Deluxe}) \\ +F_1(\text{Standard}) \end{bmatrix}$$



Source: https://en.wikipedia.org/wiki/F-score

ML Model

Performance

- **SVM** might be too simple
- **Gradient boosting** didn't overfit and had higher test score

Model	F1-macro
Dummy	0.216
SVM	0.496
Random Forest	0.542
Gradient Boost	0.548

```
--------
 Dummy
--------
Hyperparameters used are {}
The test F1-macro score is 0.21581929516985543
-----
 SVM
-----
Hyperparameters used are {'C': 0.03125}
The test F1-macro score is 0.4957812932938626
---------------
 Random Forest
---------------
Hyperparameters used are {'criterion': 'gini',
'max_depth': 80, 'max_features': 'sqrt',
'n_estimators': 10}
The test F1-macro score is 0.5420678411521804
----------------
 Gradient Boost
----------------
Hyperparameters used are {'learning_rate': 0.1,
'max_depth': 9, 'max_iter': 500}
The test F1-macro score is 0.5482298987666266
```