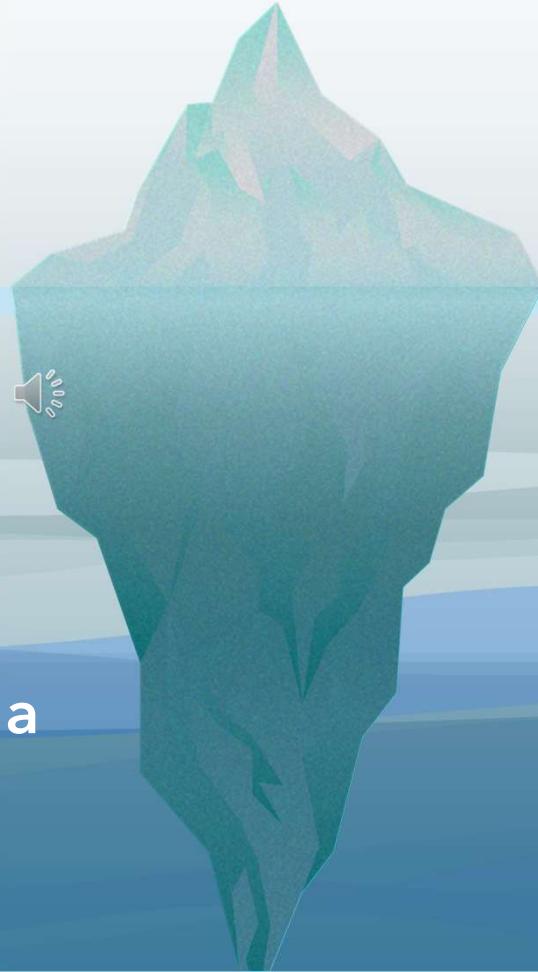


TITANIC



Presented by: Wahyu Ardhitama
Last Updated: May 13th, 2025
#Task001-P01-ML-Titanic-
20250407

Table of Contents



Introduction

Data Science and Machine
Learning Resume



About the Titanic

Goals and Objectives



Analyze

Data Wrangling and EDA



Exploratory Data Analysis

EDA – Insights and Feature
Selection



Construct

Construct and Evaluate the Model



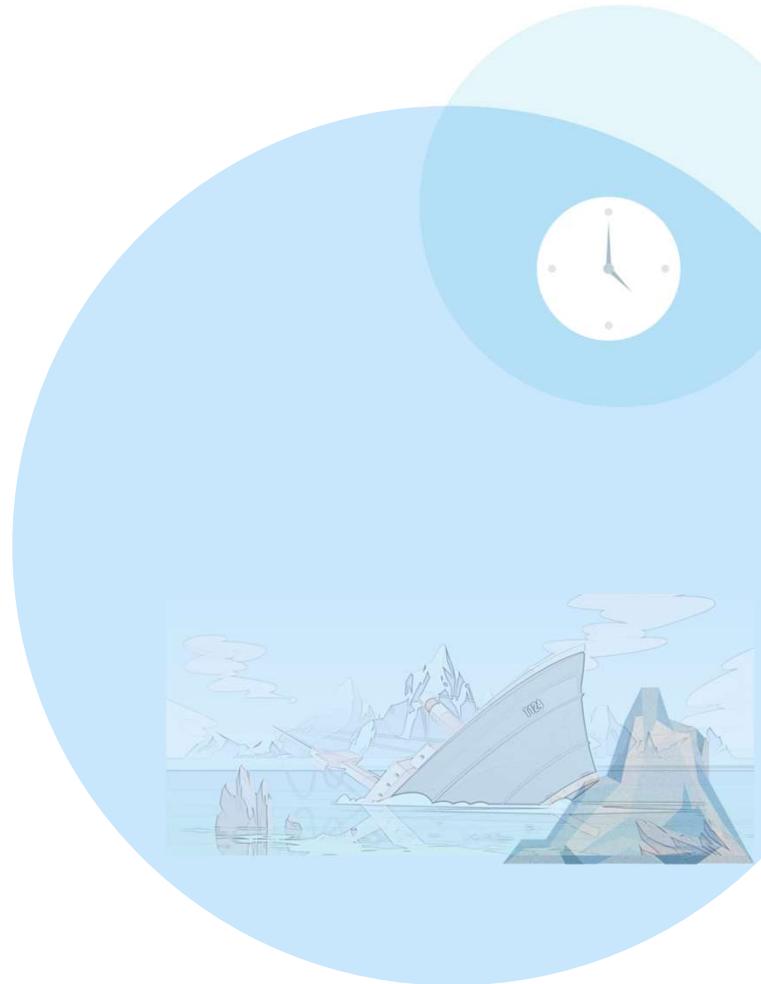
Execute

Conclusion

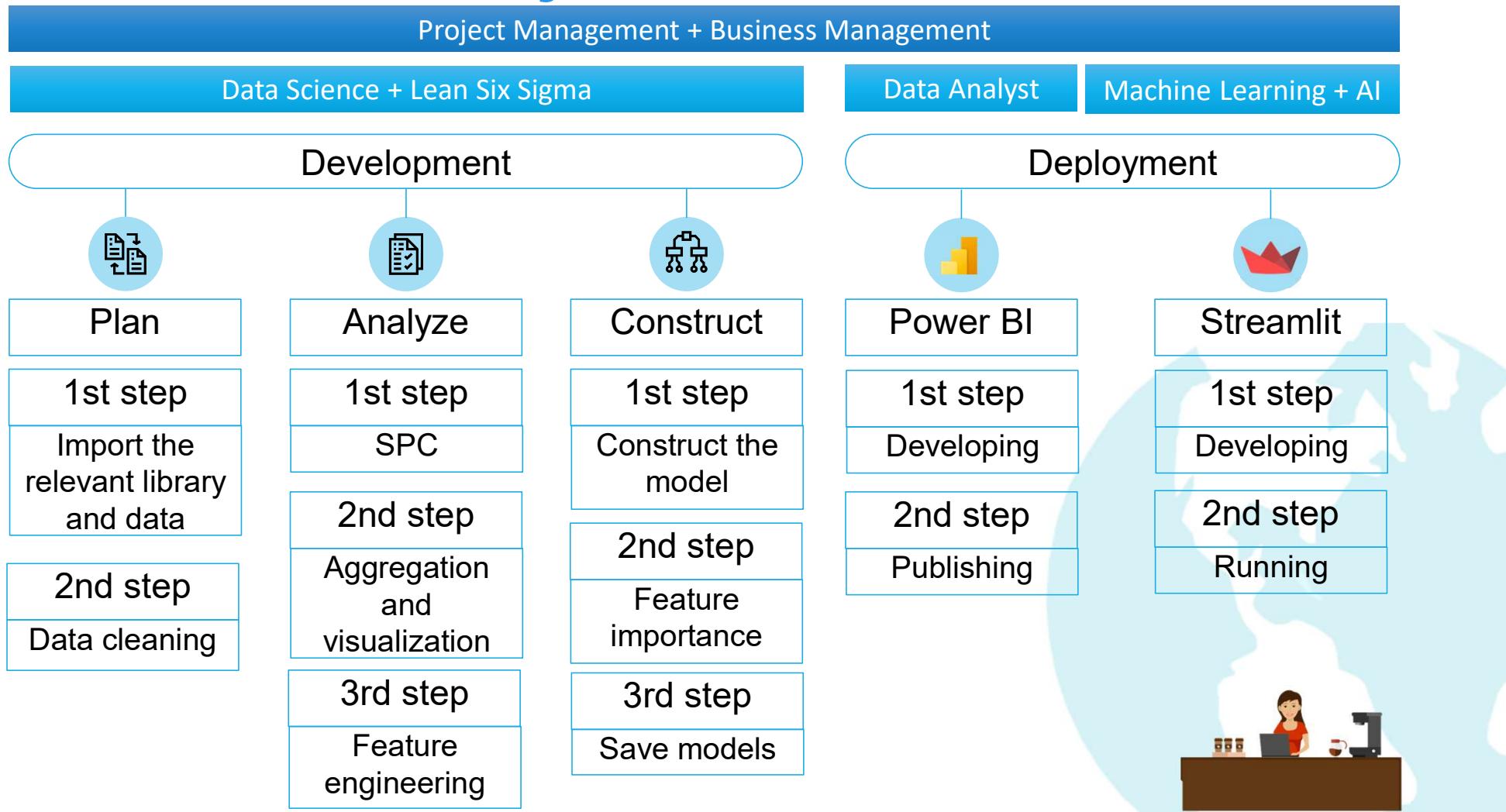
1

Introduction

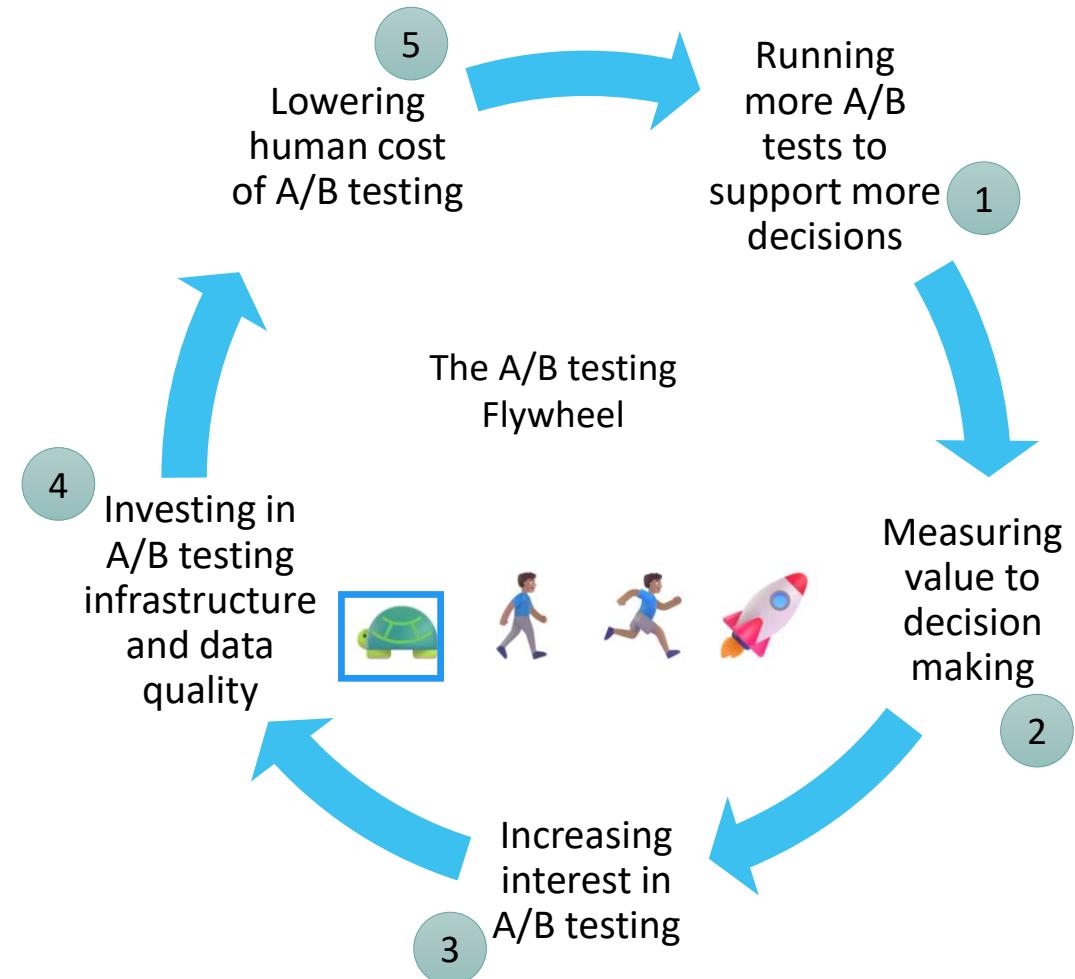
Data Science and Machine Learning
Resume



Data Project Architecture



Experimentation Life Cycle



Key takeaways :

A small or medium-sized enterprise (SME), generating monthly revenue below the 8-figure range, is currently in the early stages of growth

The organization operates with a limited budget and constrained resources, consistent with the 'Crawl' to 'Walk' phases of the business maturity model—focusing on establishing foundational processes before scaling operations

* These thresholds may vary slightly by country (e.g., EU, US, Singapore definitions)

Project Management Architecture

Project Management + Business Management

Data Analyst + Data Science + Lean Six Sigma + Machine Learning + AI

Development



Project Initiation

1. OKR and SMART
2. Stakeholder Analysis
3. Project Charter



Project Planning

1. WBS and RACI
2. Gantt Chart
3. Risk Management
4. Project Budget
5. Communication Plan

1. VOC
2. CTQ – CTP – CTR – COPQ
3. SIPOC

1. KPI Trees
2. Lean Measures
3. Operational Definitions
4. Data Collection and Sample

Deployment



Project Execution

1. Customer Survey
2. ROAM Analysis
3. Status Report
4. Hours Burndown

1. MSA
2. EDA
3. Process Capability

Agile Scrum

1. Sprint Backlog
2. Sprint Retrospective

Lean Six Sigma

1. Yellow Belt
2. Green Belt
3. Black Belt
4. Master Black Belt

1. The Process Door
2. Experimentation Platform - SPC

1. Team Meeting Agenda
2. Email management (Escalation, Invitation, Release Plan, etc.)

Revenue center
Investment center
Cost center



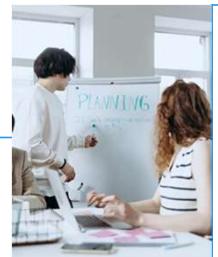
Project Online Meetings



01

Gather Consensus
Will need your feedback – Kick Off

Stakeholder requirements
Project requirements
Strategy documents
Mock up



02

Project Initiation
Prepare-discover our end goal

OKR – SMART
Stakeholder Analysis
Project Charter



03

Project Planning
This is how we're going to do

WBS and RACI
Sprint and Kanban
Project Budget
Communication Plan



04

Project Execution
Start to race

Customer Survey
ROAM Analysis
Status Report
Project Closure

Analysis and Experimentation Team

Mission:

1. Develop a platform that is easy to integrate across systems
2. Foster a data-driven culture by enabling fact-based and analytical decision-making
3. Accelerate innovation through reliable experimentation and insightful result analysis
4. Maximize ROI by aligning tasks and projects with business objectives
5. Empower stakeholders with clear, actionable, and compelling data insights

Team:

- Developers: Build the experimentation platform and the analysis tools
- **Data Scientists (AI &ML – Project/Program Managers – Data Analyst)**
- Admin

CO-V-FAST Principles: Clear/Clean/Communication/Collaboration/Correction, **O**bjective, **V**aluable, **F**ocus, **A**gile, **S**cientific and Time-bound/Trustworthiness

Projects

2025
Resume



- Web scrapping, EDA with Python and SQL
- Models: build logistic regression, Support Vector Machine (SVM), decision tree, K Nearest Neighbours (KNN) and random forest models for the prediction.

2024 Rocket Successful Landing on The First Landing



- EDA with Python
- Models: build logistic regression, Naïve-Bayes, Support Vector Machine (SVM), decision tree, K Nearest Neighbours (KNN), XGBoost, LGBoost and 5 ensemble models

2025 Kaggle: Titanic

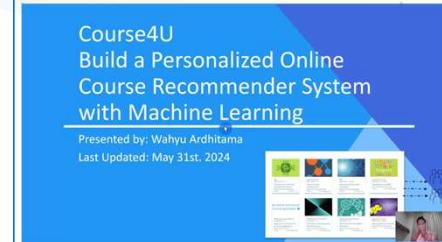
2023 NYC Citibike Business Plan

- EDA with Python, BigQuery and SQL
- Members analysis using geomap
- Models: build logistic regression, decision tree, random forest and XGBoost models to predict to increase annual membership



2024 Personal Recommender System

- Bag of Word
- Content based with user profile and course similarity (NLP)
- Clustering based and PCA
- Collaborative based with KNN and NMF, embedded features based with regression and classification
- Neural Network using Keras
- App – Streamlit
- Project Management



Wahyu
Ardhitama

Skills

2025
Resume

Wahyu
Ardhitama



Datasets

68 ↗ 2
total created



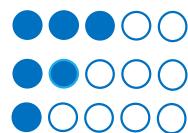
Notebooks

155 ↗ 9
total created

01

Programming Skills

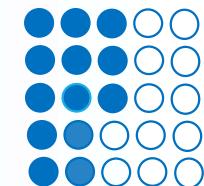
Python, SQL
Scala, R
C++, Java



02

Python Packages

Scikit_Learn, Pandas
Scipy, Seaborn
Matplotlib, Numpy
TensorFlow, Keras,
PySpark, Pytorch



03

Machine Learning Models

Supervised: Regression,
Decision Tree, Random
Forest, XGBoost, etc
Unsupervised: Kmeans,
DBScan, PCA, t-SNE



04

Deep Learning Models

LSTM, NLP
CNN, Computer Vision
LLM, RAG



2025 Projects

2025
Resume

Task001-P01-ML-ASH-20250416

Task003-P01-ML-DL-Retail-Demand-Pred-20250115

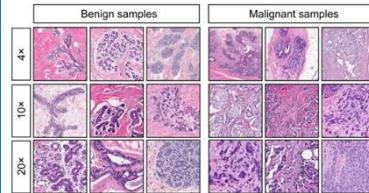


- EDA with Python
- Models: build advanced Regression, ARIMA, SARIMA, exponential smoothing Prophet and LSTM
- A/B Testing and Ad Campaign
- Power BI and Streamlit

2025 Kaggle: Sales Forecasting

2025 Wisconsin Breast Cancer

- EDA with Python
- Models: build logistic regression, Naïve-Bayes, Support Vector Machine (SVM), decision tree, K Nearest Neighbours (KNN), XGBoost, LGBoost and 5 ensemble models
- Clustering and PCA
- Neural Network using Keras and PyTorch



Task001-P01-DS-AB-Testing-Ad-TV-Campaign-20240131



- EDA with Python
- Advanced regression, random forest and gradient boosting
- Clustering and t-SNE
- Neural Network using Keras

2025 Kaggle: Titanic

Wahyu
Ardhitama

2025 Computer Vision: MNIST Fashion

- EDA with Python
- Clustering and PCA
- Neural Network using Keras and PyTorch

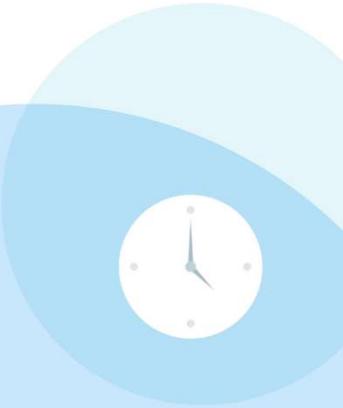
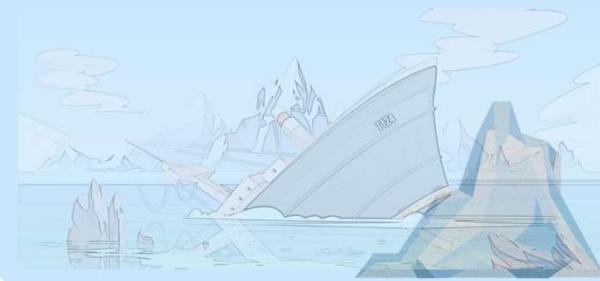


2025 Recommender System with LLM MAB Bayesian Optimization

2

About the Project

Business Case and Objective



ABOUT THE TITANIC

Top
5%



Titanic - Machine Learning fro...
0 Submissions left today
Ongoing · 777/16248

Executive Summary

890
people

Train data

420
people

Test Data

78.9%
1st Submission

79.9%
Best Submission

12
ML Models

82%
Metric



Analytics

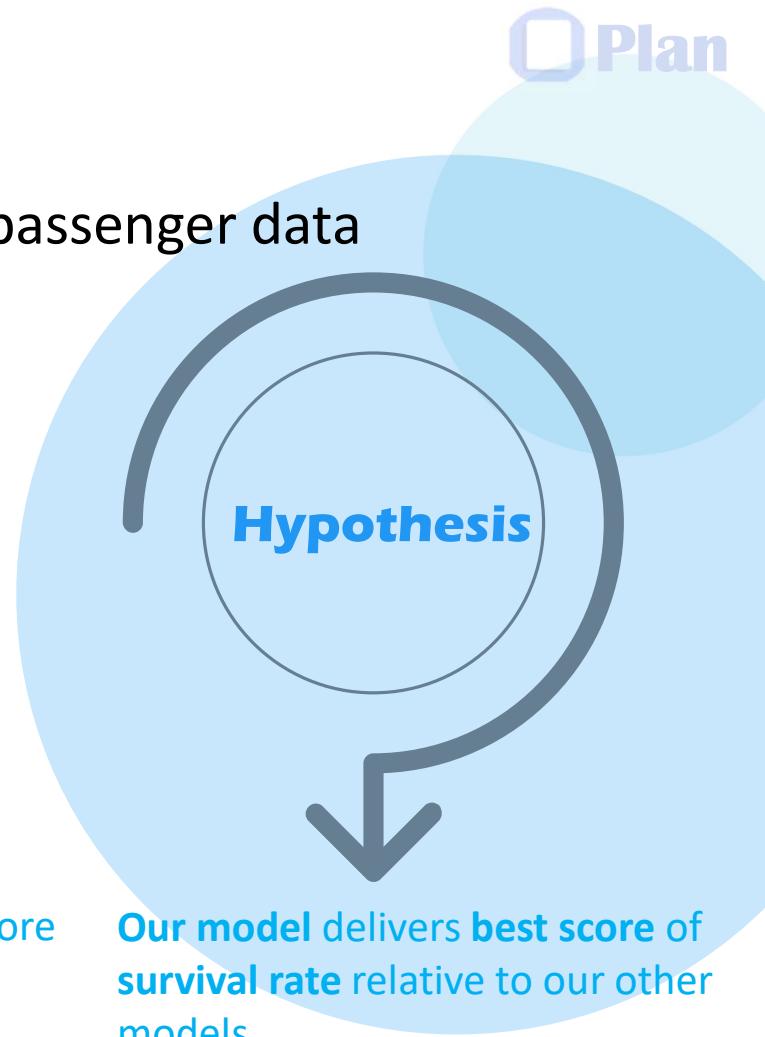
Predicting Titanic survivors based on passenger data



Explore and compare **various machine learning models** and **find one** with the **best performance** to predict Titanic survivors based on passenger data



- Acquire **more than 79% score** from Kaggle **k**
- Deliver **valuable insights**
- Submit and compare **results**



Our model delivers best score of survival rate relative to our other models

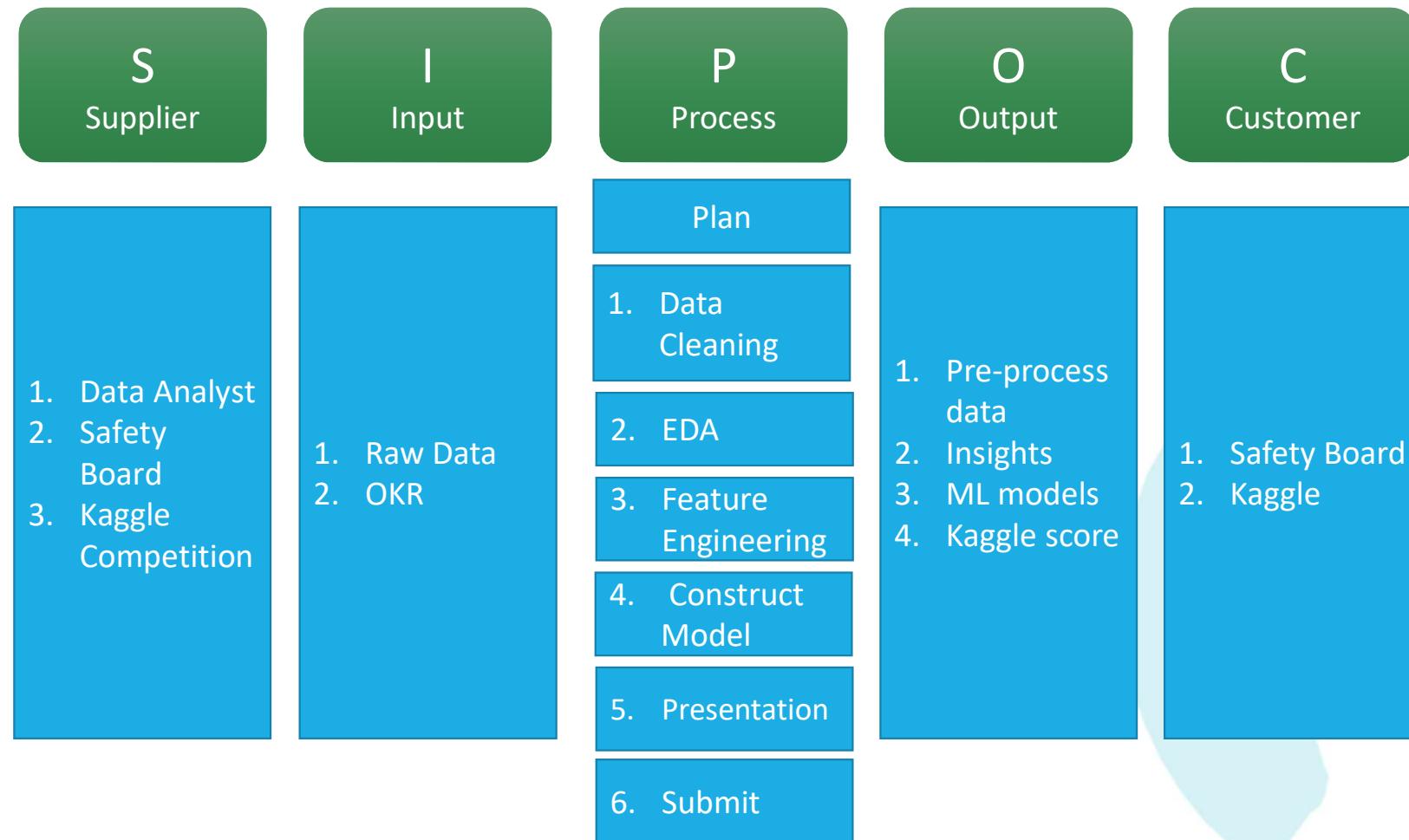
Plan

Task Charter

Problem & solution	Based on Titanic data, fewer than 40% survived . Developed a machine learning model predicting survival probability for safety protocols and future maritime endeavors.
Methodology	<input type="checkbox"/> PACE framework <input type="checkbox"/> Agile project management <input type="checkbox"/> Lean 7 Waste + SPC + Six Sigma
Goals & Improvement metric	<input type="radio"/> Primary : More than 79% score <input type="radio"/> Guardrail : 82% accuracy; 80% precision

Titanic Competition

SIPOC



Process Waste

The 7 Wastes Plus

7 Waste + In Lean (Muda)	Findings	Mitigate Strategy
Waiting	Long training time for complex models (e.g., LGBM, XGBoost, Gradient Boosting, Voting Classifier)	Use faster models (e.g., Random Forest), subset data, or use GPU acceleration and parallel processing
Inventory	Storing too many intermediate models, datasets, or unused feature sets	Set clear file naming/versioning standards and delete obsolete models/datasets regularly.
Transport	Moving data between Kaggle, Colab, local disk, etc.	Centralize work in one environment (e.g., Kaggle Notebooks) and automate data sync using scripts
Motion	Manually switching between environments or rerunning unchanged notebooks	Use automated pipelines (e.g., Snakemake, MLflow) and notebook checkpoints to avoid redundant steps
Overproduction	Building many models before validating data or establishing baselines	Start with a baseline (e.g., Logistic Regression), validate EDA findings, then scale up model complexity
Defects	Data leakage, mislabeled data, or flawed preprocessing causes inaccurate models	Perform thorough EDA and validation; use cross-validation; implement data quality checks; research
Extra Processing	Using ensemble models when simpler ones (e.g., Random Forest) achieve similar accuracy	Benchmark model complexity vs. accuracy early; favor interpretable, simpler models when performance is close
Unused Talent	Not involving teammates in feature engineering, domain insights, or result interpretation	Promote collaborative reviews, assign roles (e.g., feature lead, model validator), encourage diverse input

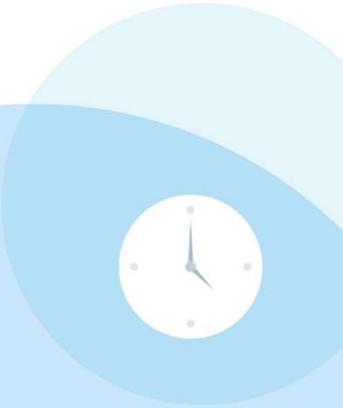
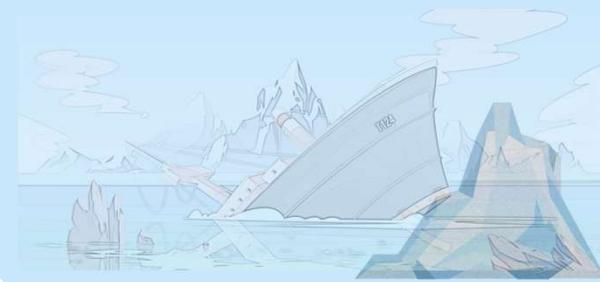
OKR Status

Objective	Develop a High-Performing Survival Prediction Model	Benchmark and Evaluate Multiple ML Models	Enhance Feature Engineering and Data Quality
Status	In Progress	Next	Next
Progress Score	<div style="width: 60%;"></div>	<div style="width: 60%;"></div>	<div style="width: 80%;"></div>
Deadline	May 1, 2025	May 5, 2025	May 7, 2025
Key Result 1	Achieve Kaggle submission score >75% ✓	Train the promising machine learning models ✓	Reduce missing data to <1% across all relevant features ✓
Key Result 2	Reach validation accuracy, precision, recall, and F1-score >80% ✓	<ul style="list-style-type: none">▪ Random forest▪ Gradient boosting▪ Logistic regression	Create at least 5 new meaningful engineered features ✓
Key Result 3	Final model has generalization gap <5% between training and validation accuracy ✓	Reduce runtime for model training by 20%	Improve model performance by at least 5% after feature engineering ✓

3

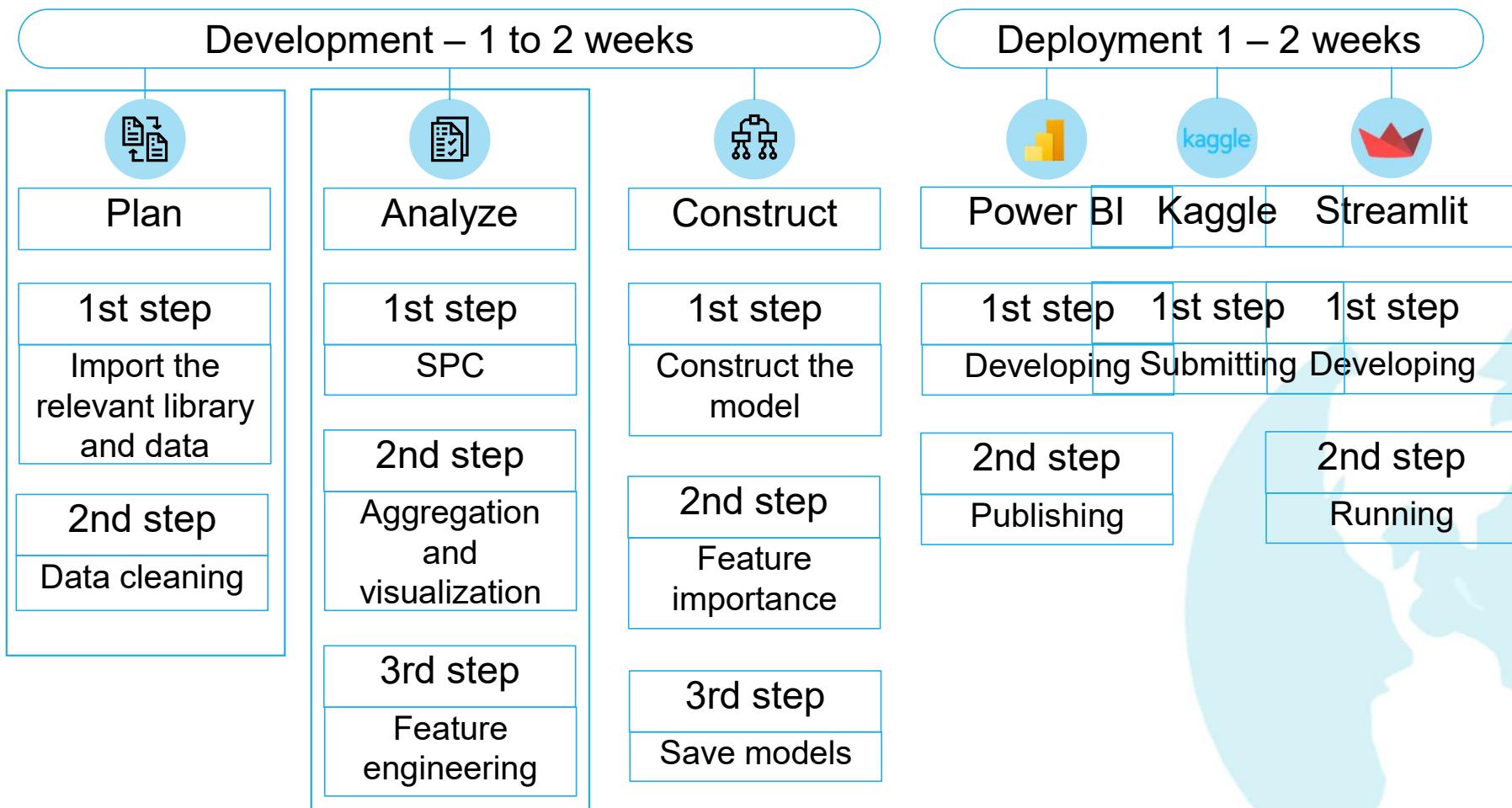
Analyze

Data Wrangling and EDA



Data Project Architecture

Plan



Titanic Data

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	train_set	Survived
1171	2	Oxenham, Mr. Percy Thomas	male	22.0	0	0	W./C. 14260	10.5000	NaN	S	0	NaN
962	3	Mulvihill, Miss. Bertha E	female	24.0	0	0	382653	7.7500	NaN	Q	0	NaN
976	2	Lamb, Mr. John Joseph	male	NaN	0	0	240261	10.7083	NaN	Q	0	NaN
928	3	Roth, Miss. Sarah A	female	NaN	0	0	342712	8.0500	NaN	S	0	NaN
980	3	O'Donoghue, Ms. Bridget	female	NaN	0	0	364856	7.7500	NaN	Q	0	NaN
1306	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9000	C105	C	0	NaN
1021	3	Petersen, Mr. Marius	male	24.0	0	0	342441	8.0500	NaN	S	0	NaN
939	3	Shaughnessy, Mr. Patrick	male	NaN	0	0	370374	7.7500	NaN	Q	0	NaN
894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q	0	NaN
1233	3	Lundstrom, Mr. Thure Edvin	male	32.0	0	0	350403	7.5792	NaN	S	0	NaN

all_data_df

train_df (test.csv)

test_df (test.csv)

10 Features (X)

Numerical

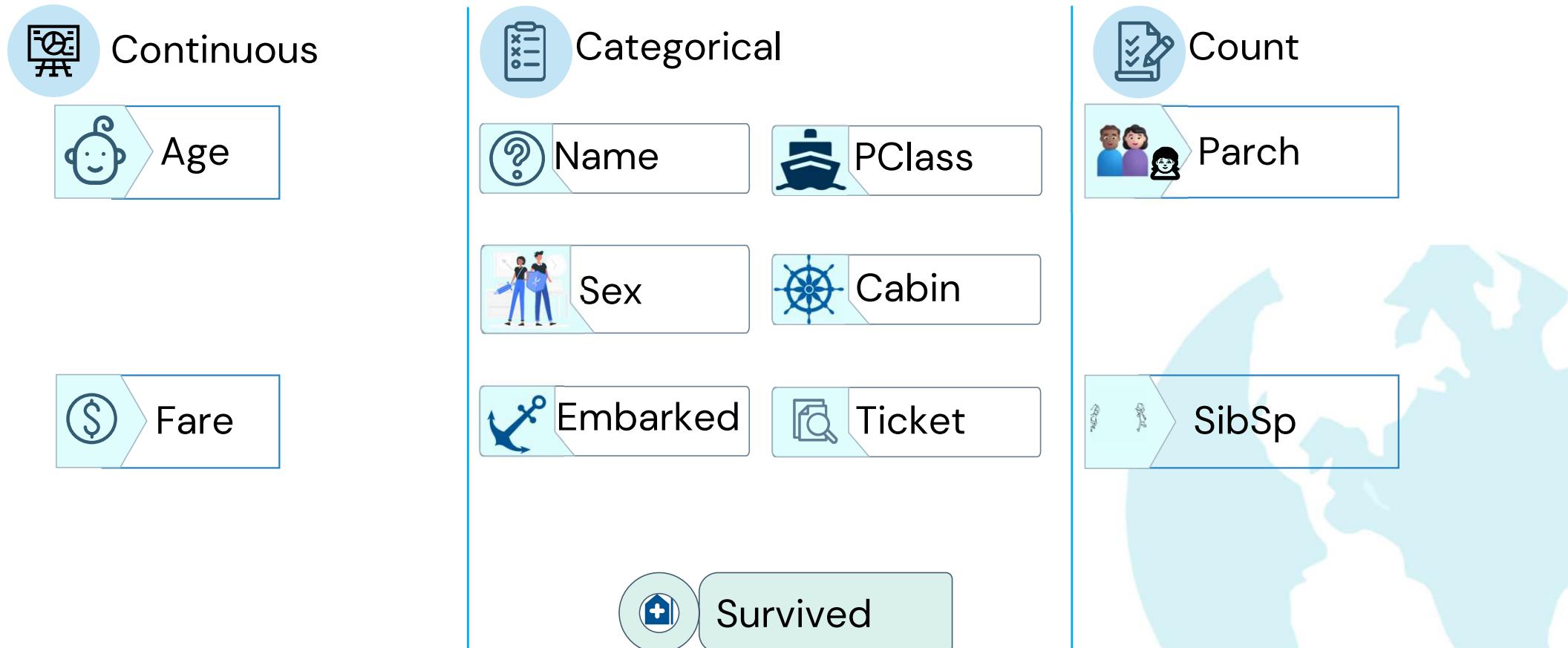
Categorical

Count

Target

Binary Label

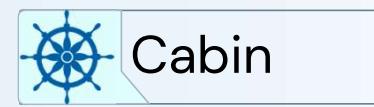
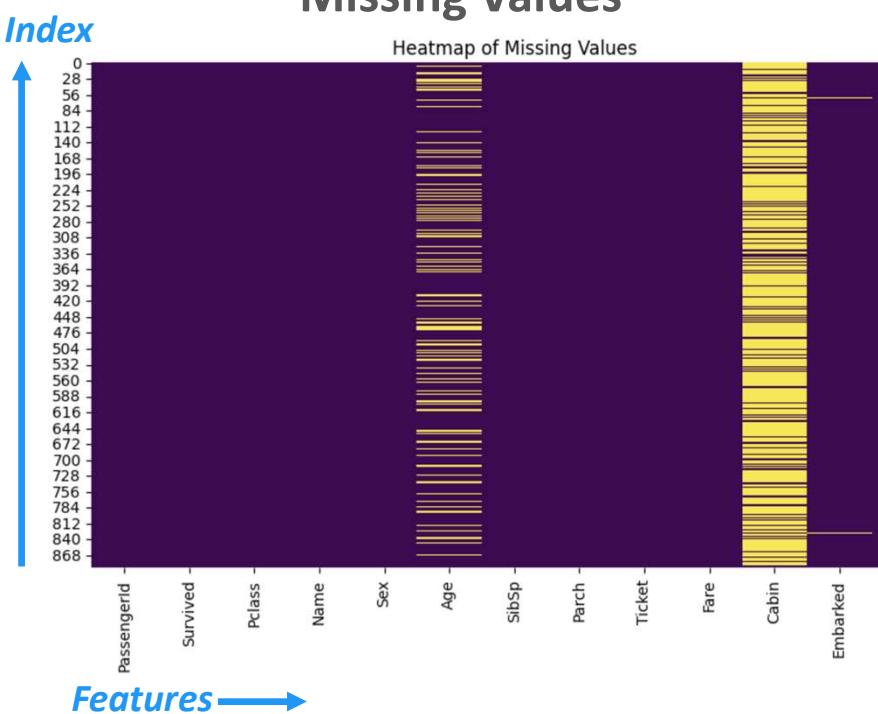
Titanic Data



Titanic Data

Data Cleaning

Missing Values



Proportional –
CumSum < 51%



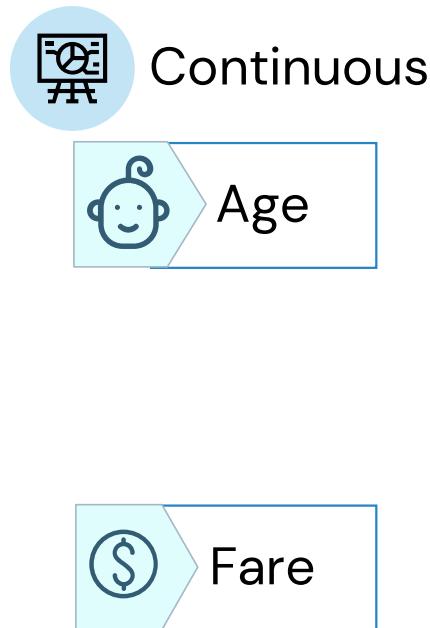
U

O

0 - 4

to the highest

Titanic Data



Data Wrangling



Categorical



Name



PClass



Sex



CabinGroup



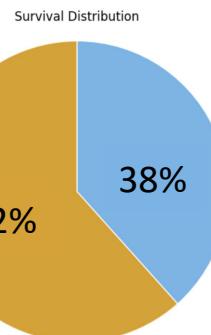
Embarked



TicketNo



Survived



Count



Parch



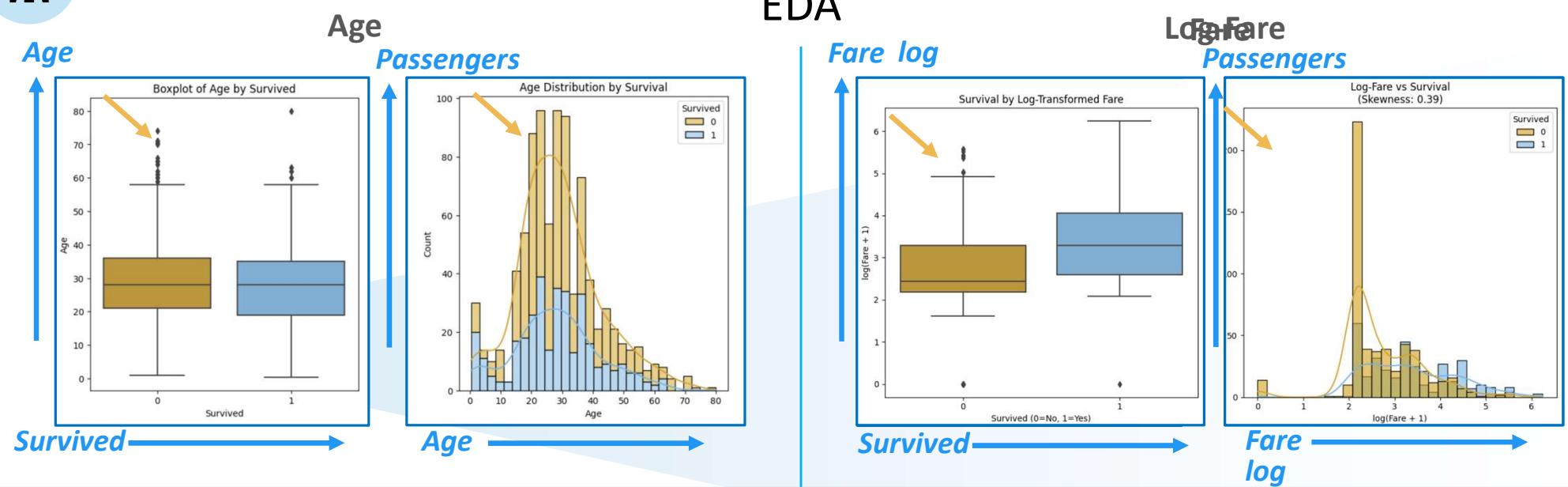
CabinGmt

Key takeaways

Survived passengers = 340 (38%)
Deceased passengers = 550 (62%)

Age and Fare

EDA

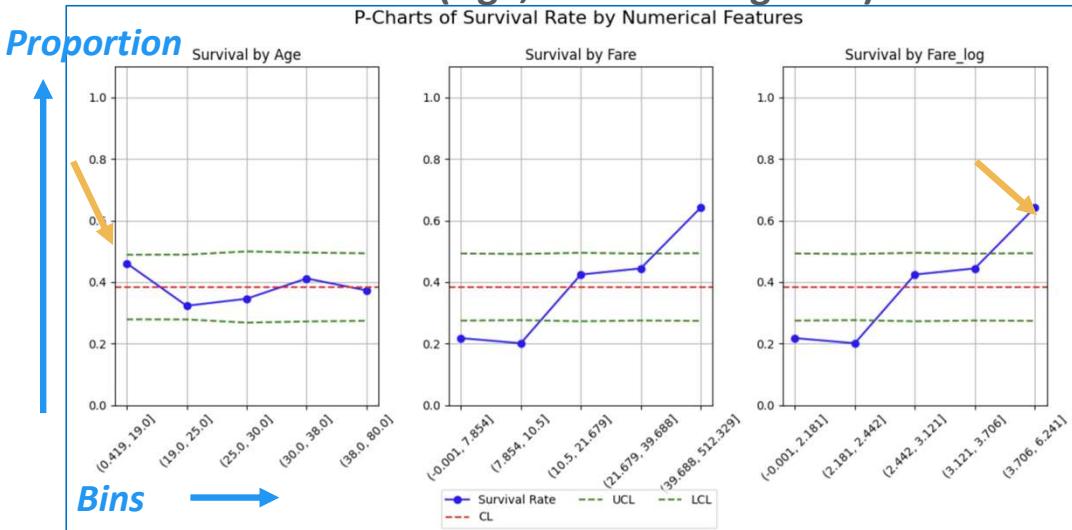


Practical Interpretations:

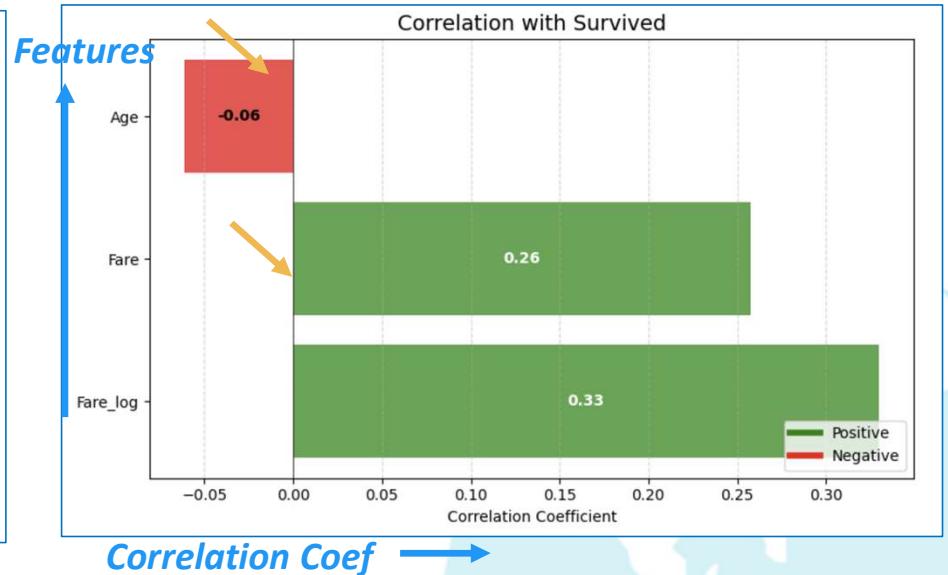
Variable	Normality Conclusion	Skewness Interpretation	Kurtosis Interpretation	Recommendation
0 Age	Non-normal	Moderately skewed (0.54)	Light tails (0.72)	Consider non-parametric tests
1 Fare	Non-normal	Highly skewed (4.79)	Heavy tails (outliers likely) (33.40)	Try log/Box-Cox transform + outlier treatment
2 Fare_log	Non-normal	Symmetric (0.39)	Light tails (0.98)	Consider non-parametric tests

Age and Log-Fare Insights

P-Charts (Age, Fare and Log-Fare)



Numerical Features Correlation



Key takeaways & solutions:

Passengers **under 19 years** old showed **higher survival rates**.

Select **log(fare)** over raw fare values due to its stronger correlation

Higher-paying passengers had significantly better survival chances.

Revenue per survived passengers = \$48 (38%)

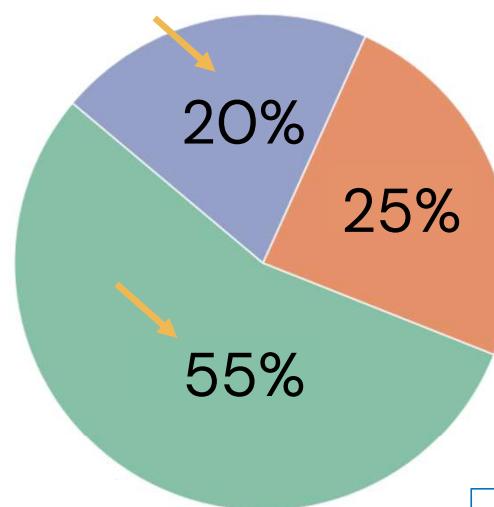
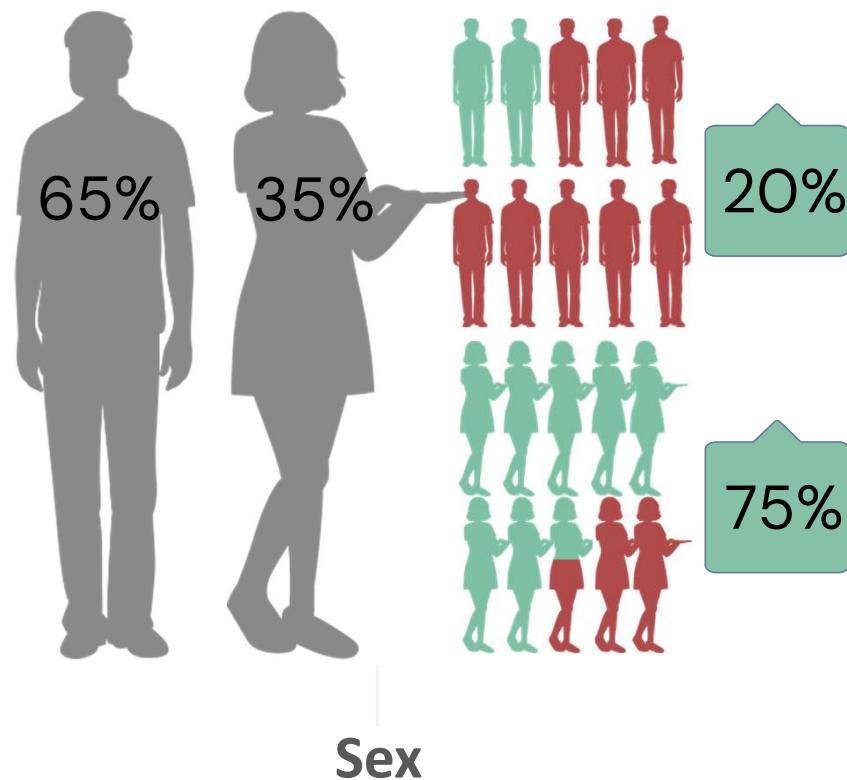
Revenue per deceased passengers = \$22 (62%)

Next: Elaborate 'Age' into child (under 13) adolescent (13 – 19), adult (19-64) and seniors

Sex and Pclass Insights

Key takeaways:

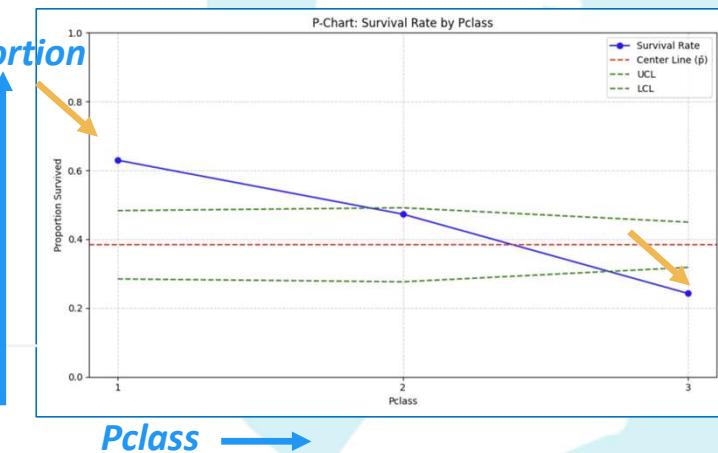
A higher proportion of females survived



Key takeaways:

3rd Class had the highest passengers
2nd class had the lowest

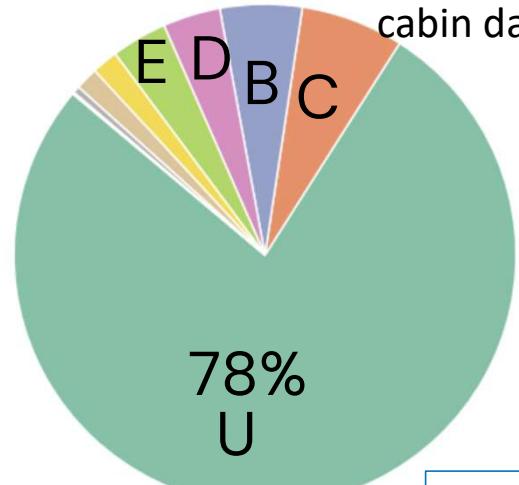
1st Class had better survival chance
3rd class had the lowest



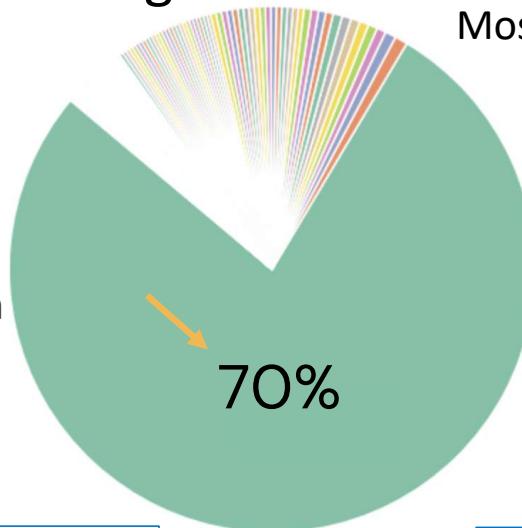
Cabin Group and Cabin No

Key takeaways:

Most passengers fall into the group with unknown cabin data

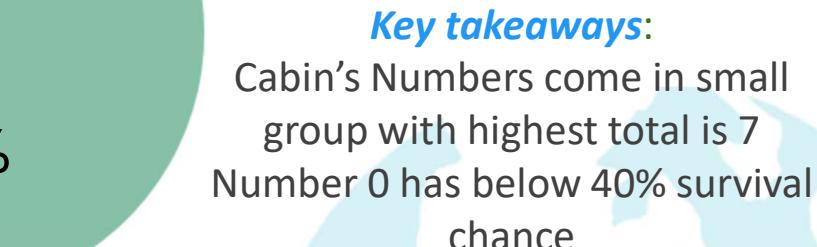


Insights



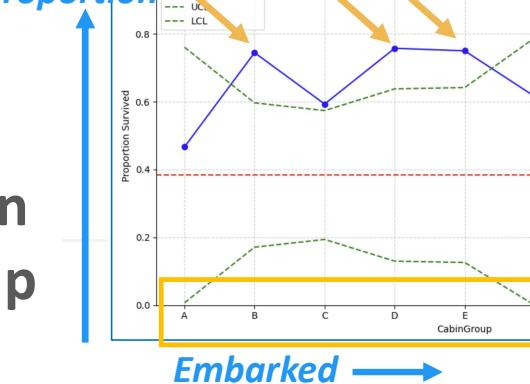
Key takeaways:

Most passengers Cabin's Number is 0, unknown Cabin



Proportion

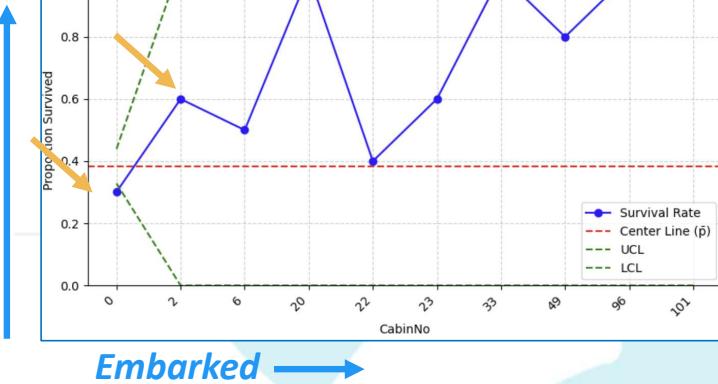
Cabin Group



Embarked →

Proportion

Cabin No

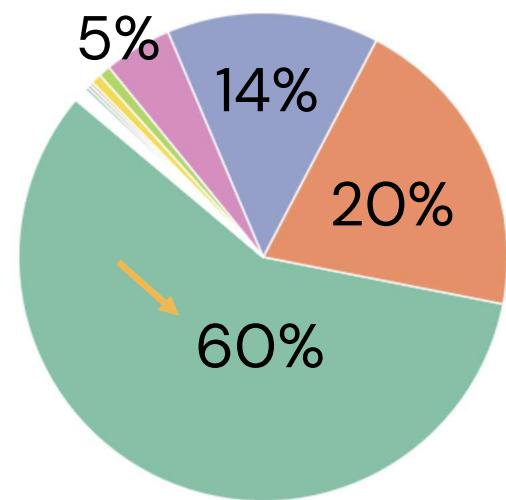


Embarked →

Title

Key takeaways:

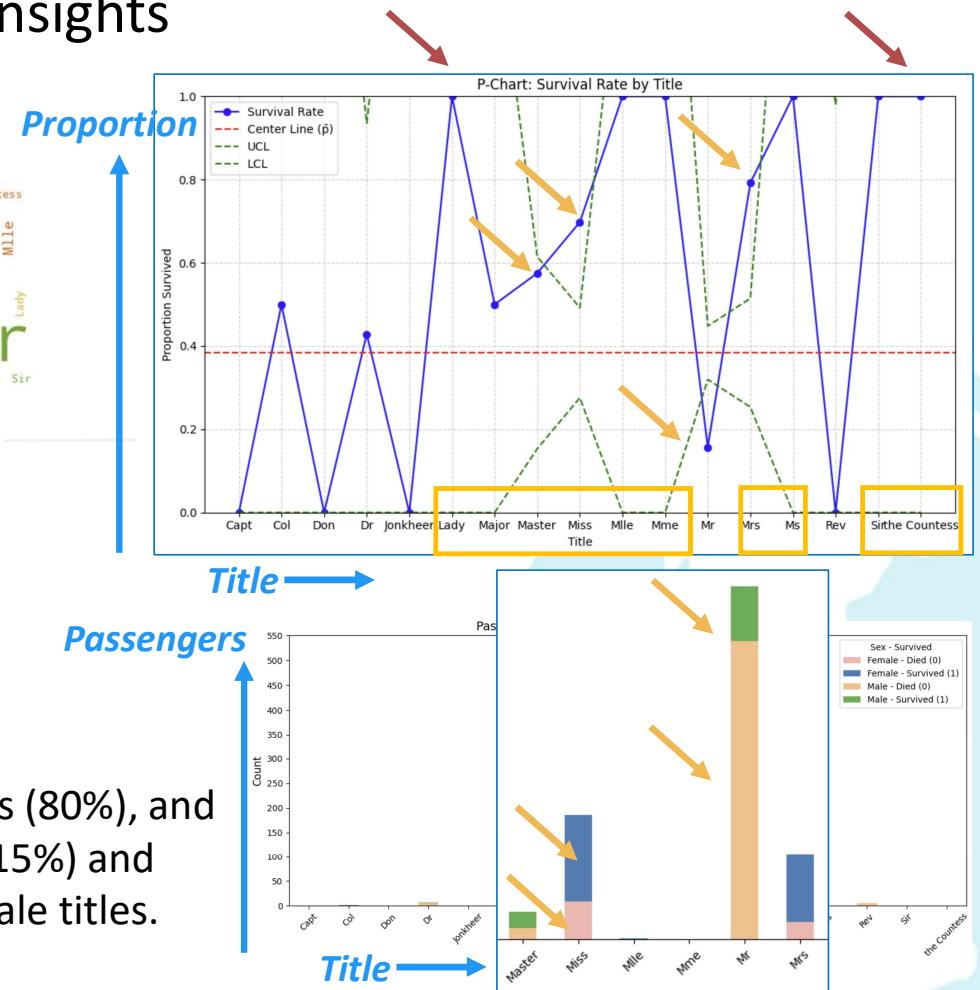
The majority of passengers hold the title 'Mr', followed by 'Miss', 'Mrs', and 'Master'



Key takeaways:

Female passengers with titles such as Miss (70%), Mrs (80%), and noble titles had higher survival rates, while 'Mr' (15%) and 'Master' (55%) were among the most common male titles.

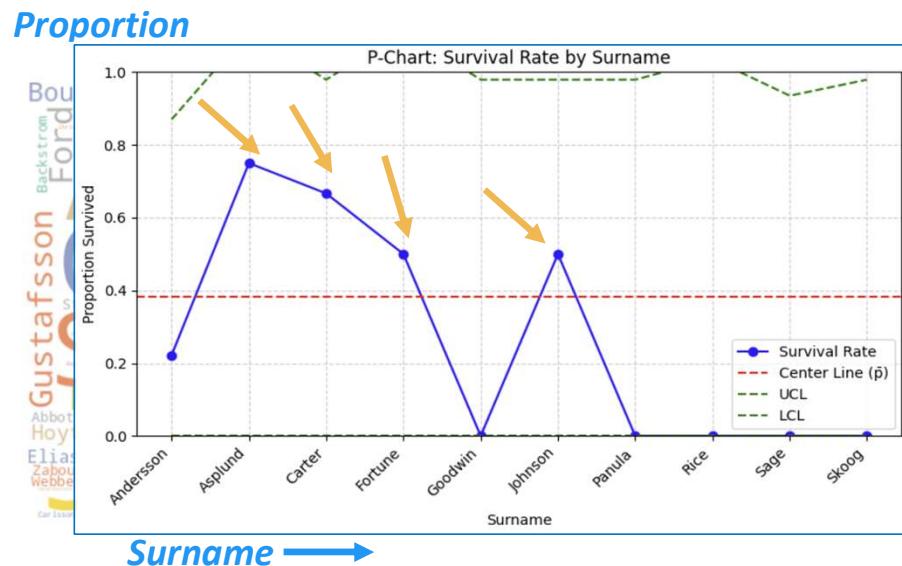
Insights



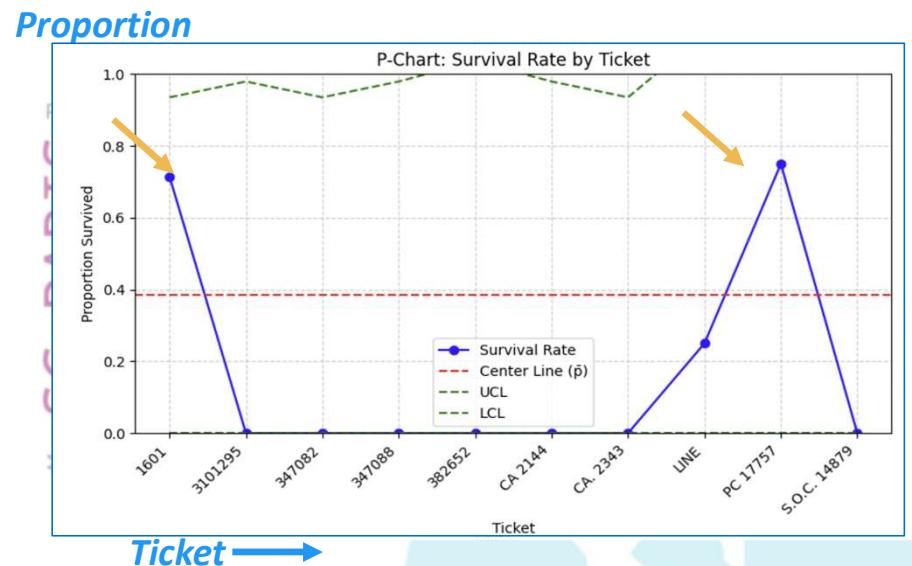
Surname and Ticket

Insights

Surname



Ticket



Key takeaways:

Displayed as is, we can see that both 'Surname' and 'Ticket' contain many unique values (variables). We retain them not for interpretability, but to preserve potential patterns and signals in the data

Passengers with surnames such as **Asplund, Carter, Fortune, and Johnson** appear to have had **better chances of survival**. Anderson appearing in up to 9 members with only 2 survived.

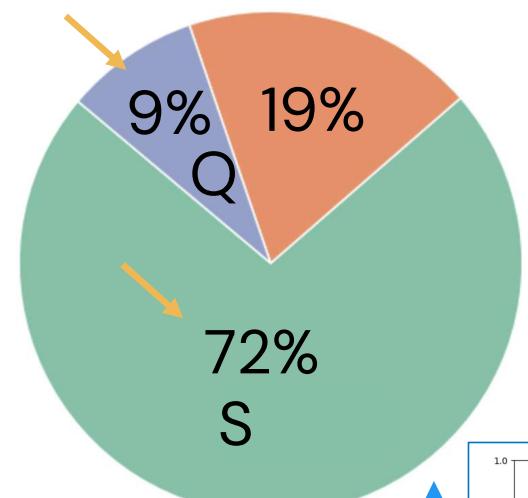
Similarly, certain **ticket numbers** are associated with **higher survival rates**.

Embarked and Multiple Cabin



Key takeaways:

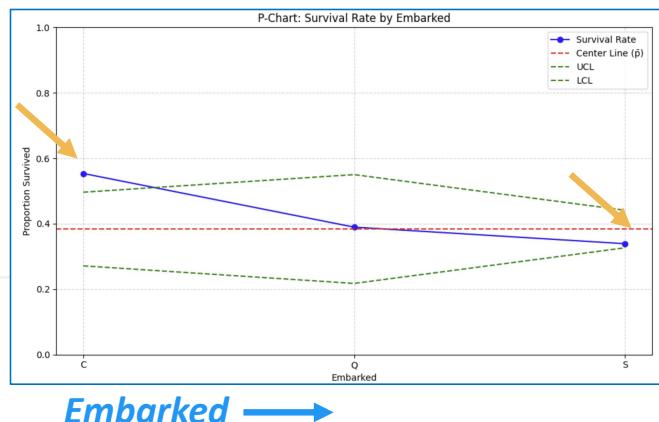
Most passengers embarked from port S
Port Q had the lowest number of departures



Proportion

Embarked

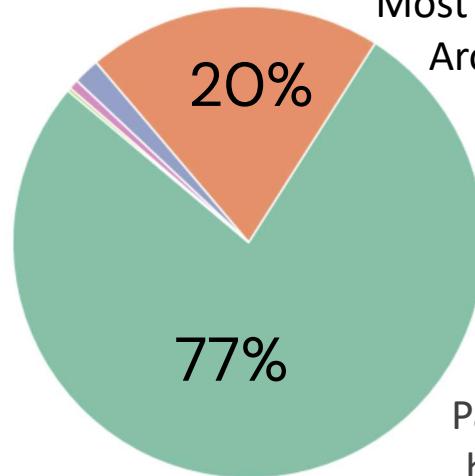
Key takeaways:
Passengers from port C had a better survival chance
Port S had lower chance



Insights

Key takeaways:

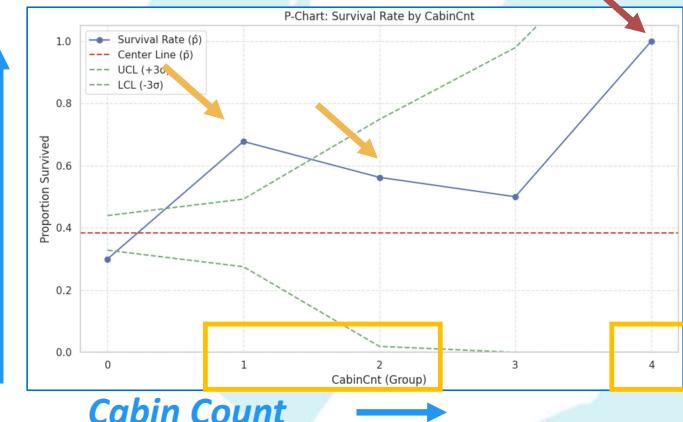
Most passengers had no cabin
Around 20% had 1 cabin



Proportion

Cabin Count

Key takeaways:
Passengers who had 1 cabin had better survival chance

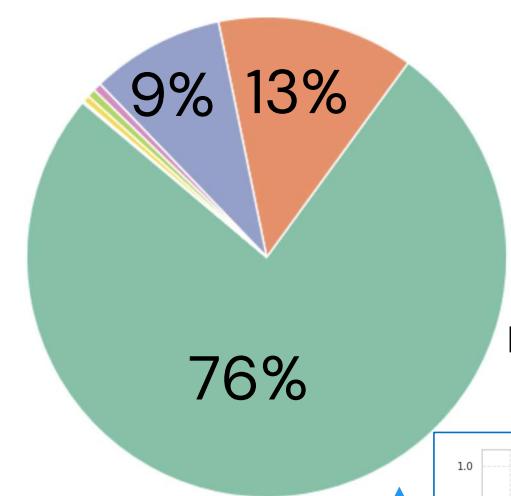


Parents/children and Sibling/Spouse

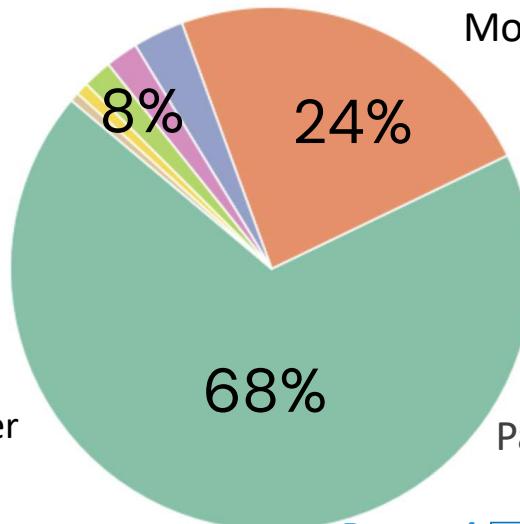
Analyze

Key takeaways:

Most passengers traveled alone (solo), followed by those with 1-2 parents/children (Parch).



Insights



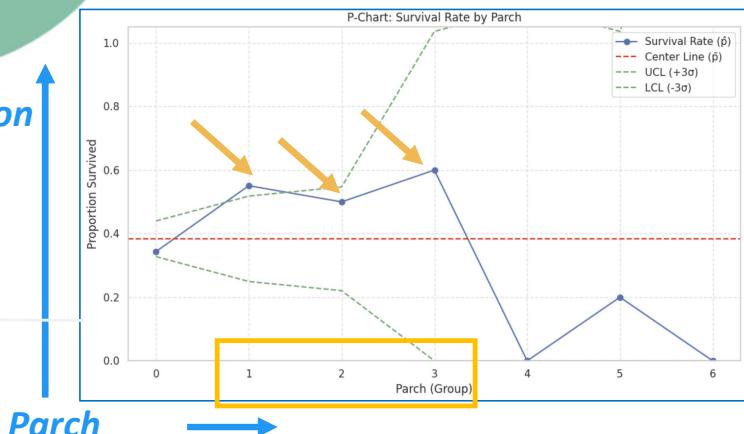
Key takeaways:

Most passengers traveled alone (solo), followed by those with 1-2 siblings/spouse (SibSp).

Key takeaways:
Passengers with 1-3 parents/children had higher survival rates

Proportion

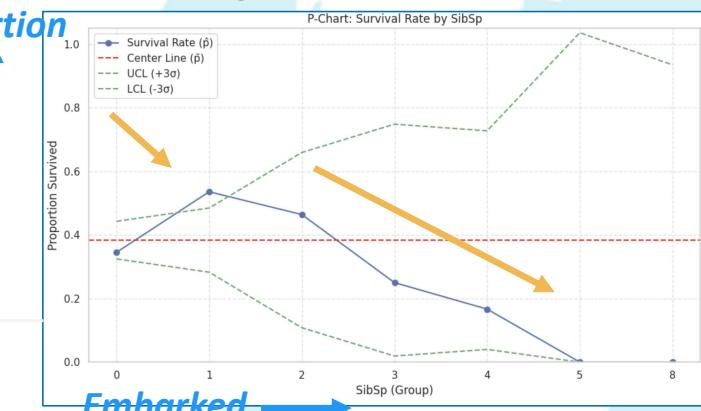
Cabin Group



Embarked

Proportion

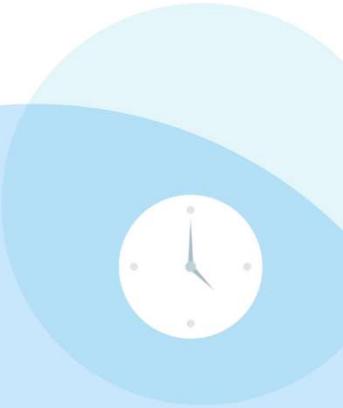
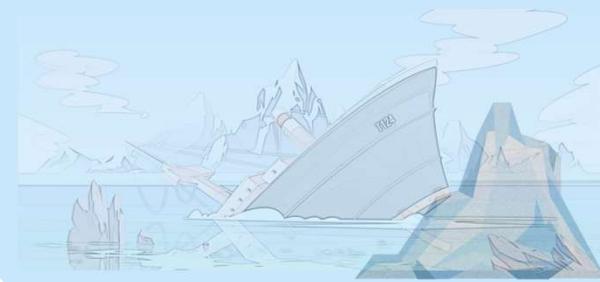
Key takeaways:
Passengers with 1 – 2 siblings/spouse had higher chance of survival



4

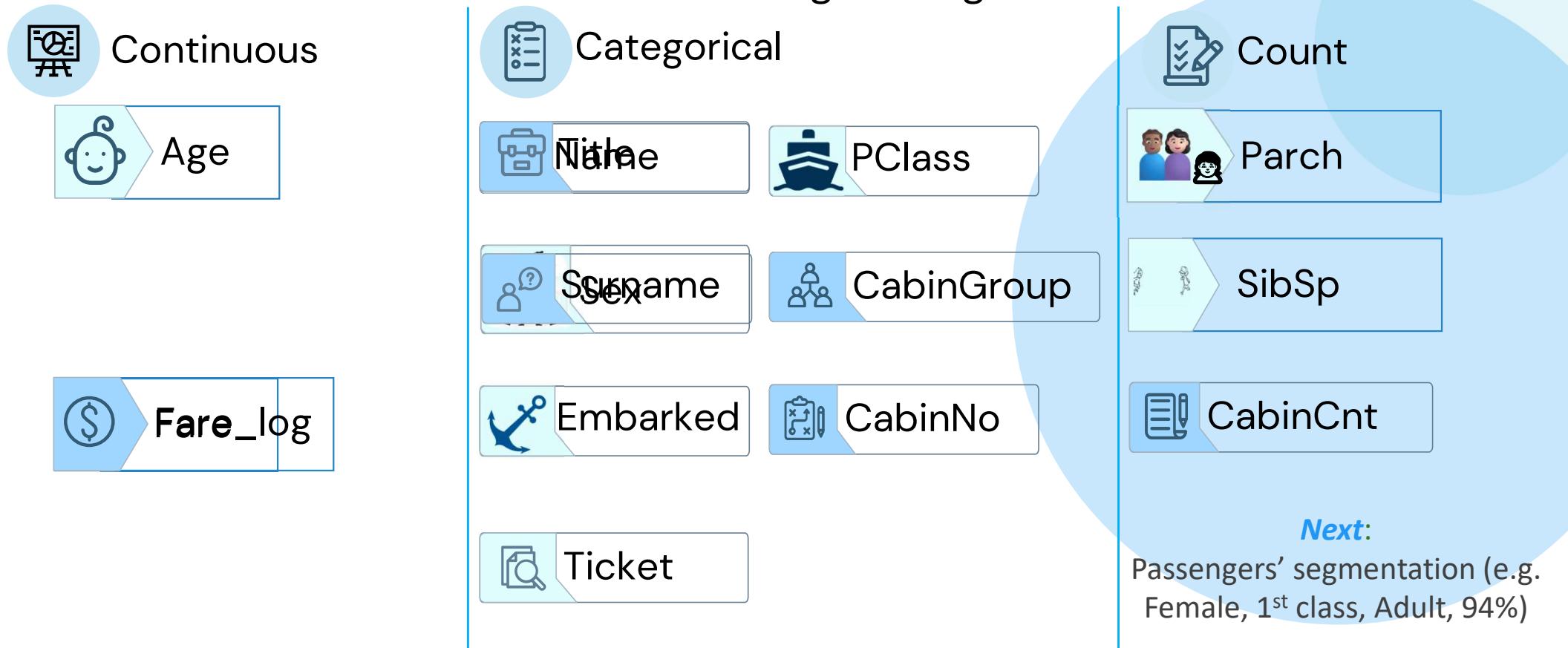
Exploratory Data Analysis

EDA – Insights and Feature Selection

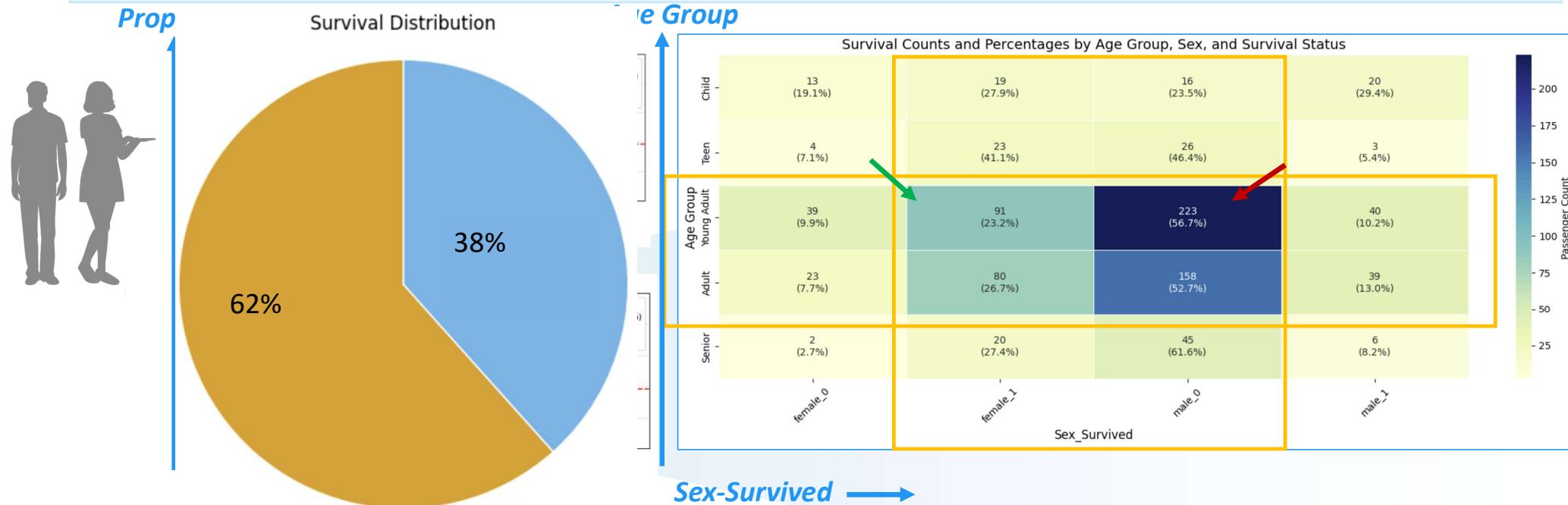


Titanic Data

Feature Engineering



Age Group, Sex and Survived



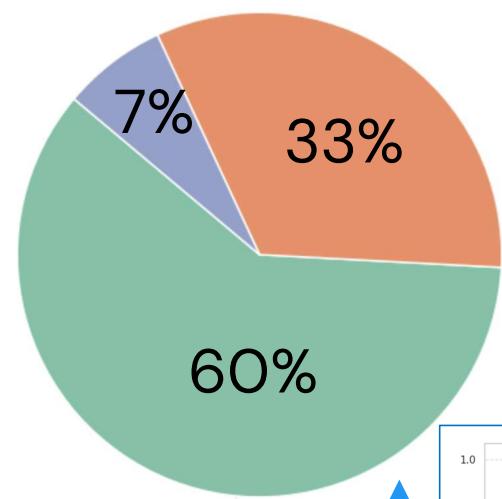
Age Group	Male SR	Female SR	Key Observations
Children (~70)	~30% 🧑	~30% 🧑‍🦰	Highest male survival rate among all age groups
Teen (~55)	~5% 🧑	~40% 🧑‍🦰	Teenage females had near-highest survival; males fared worse than children.
Young Adult (~390)	~10% 🧑	~25%	Males faced highest fatalities (>85%); female survival remained strong
Adult (300)	~15 🧑	~25%	Severe male mortality (>80% died); females still prioritized
Senior (~70)	~10% 🧑	~30%	Highest male fatalities rate among all age groups

Family Group and Family Size

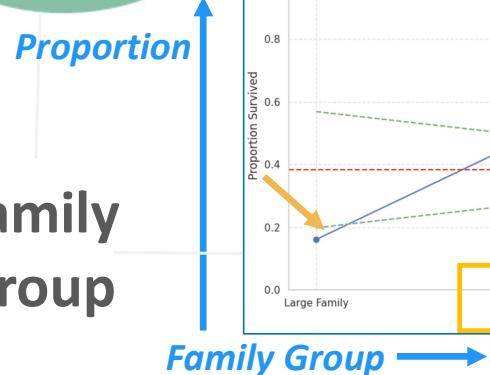


Key takeaways:

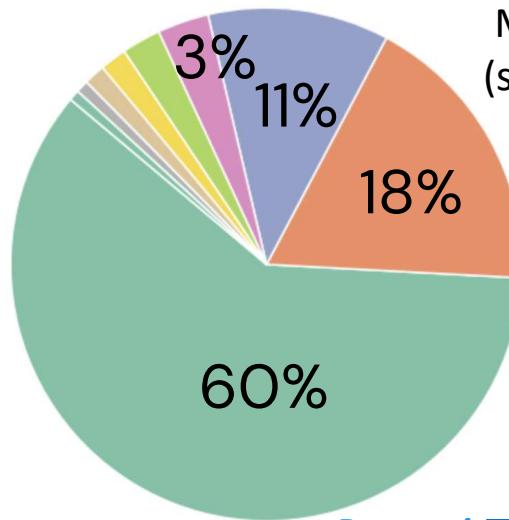
Most passengers traveled alone (solo), followed by those with small family



Key takeaways:
Small family had higher survival rates

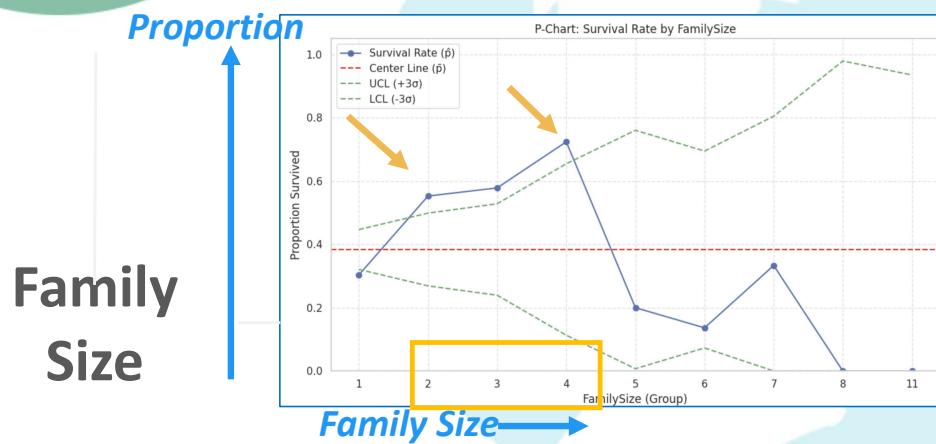


Insights



Key takeaways:

Most passengers with 1 family size (solo), followed by those with 2 to 4 sizes



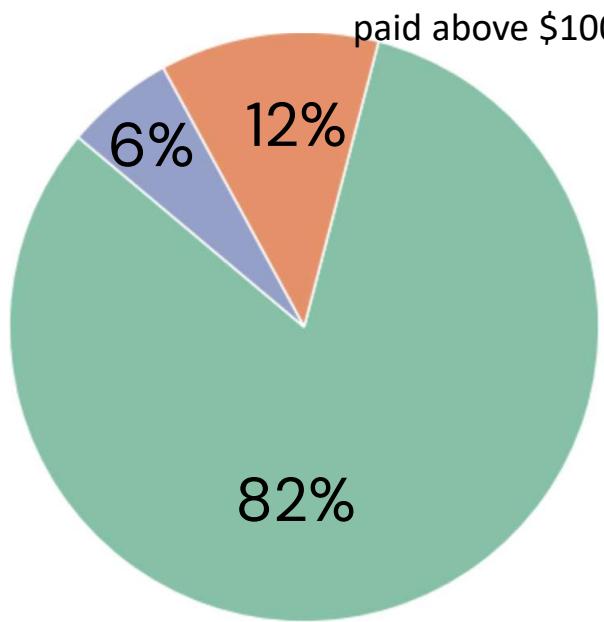
Family Group

Family Size

Fare Group

Key takeaways:

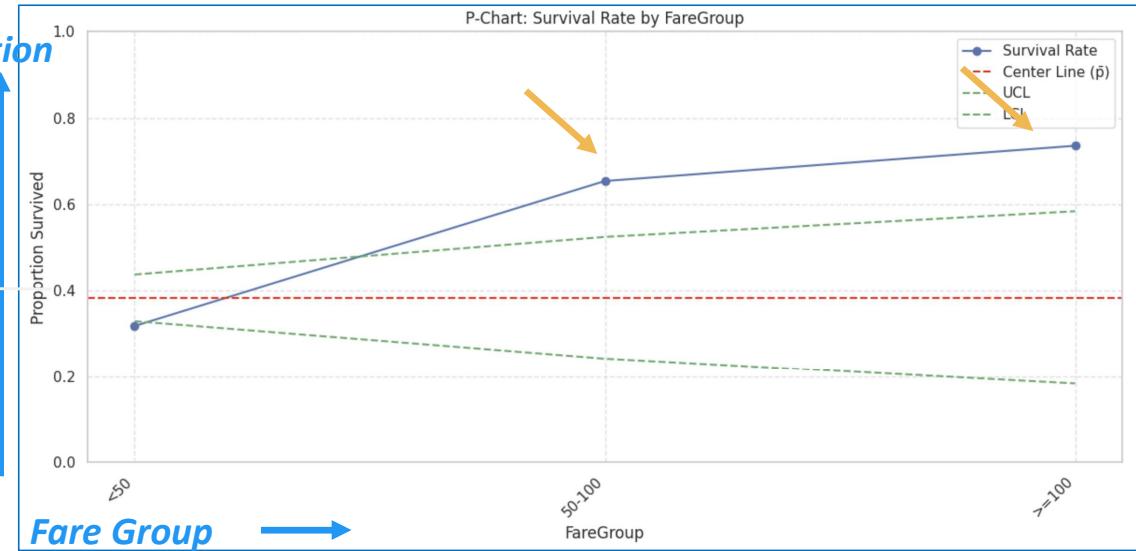
Most passengers paid below \$50, followed by those who paid between \$50 to \$100, and then those who paid above \$100



Fare Group

Insights

Proportion

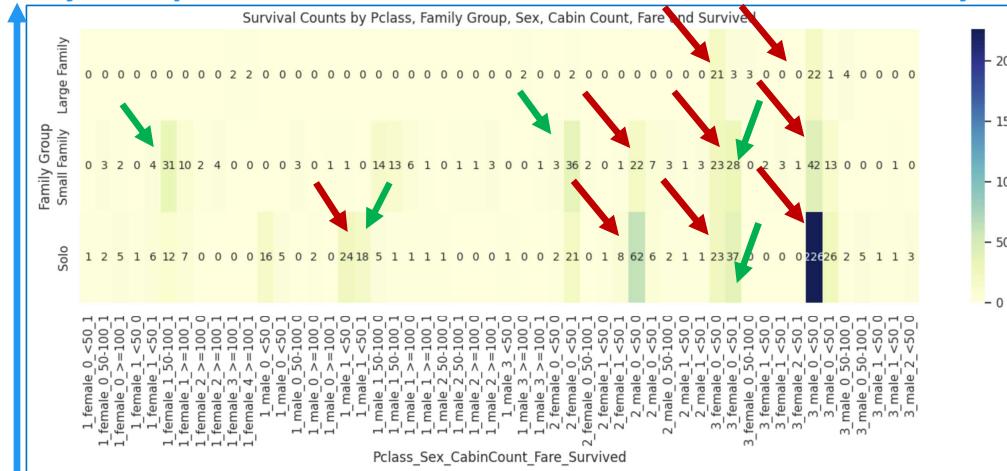


Key takeaways:

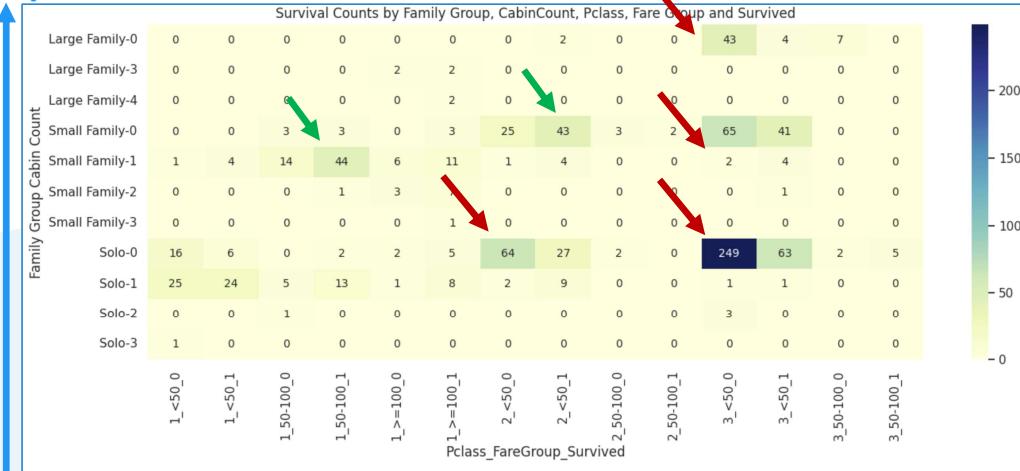
Passengers with higher-fare tickets (>\$100) had the highest survival rate (~75%) representing only 6% of all passengers

Family Group, Pclass, Cabin Count, Fare and Survived

Family Group



Family Group – Cabin Count



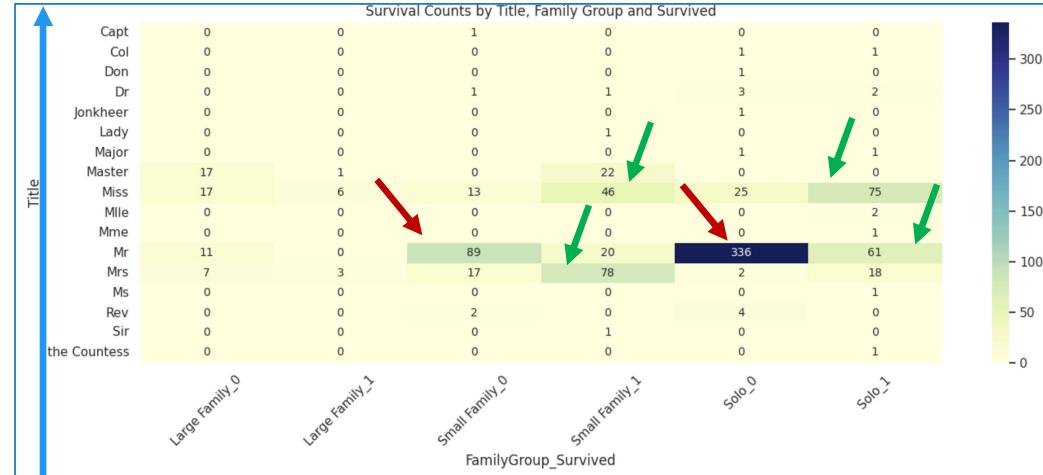
Pclass-Sex-Cabin Count-Fare- Survived →

Pclass-Fare- Survived →

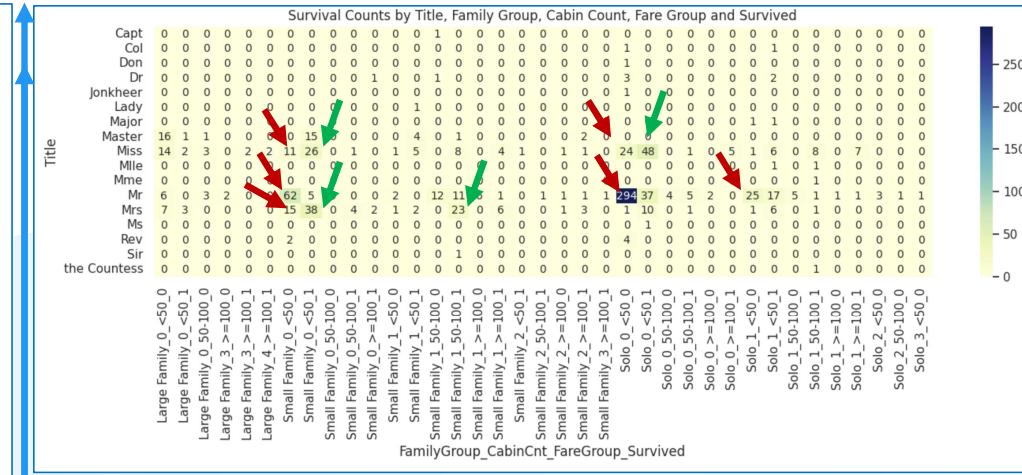
Family Group	Class	Cabin Count	Fare	Survival	Key Observations
Solo	3 rd , 2 nd	0	< \$50	Mostly deceased	Highest fatality rate and second-highest male fatality
Small Family	3 rd , 2 nd	0	< \$50	Mostly deceased	Severe male fatality group
Large Family	3rd	0	< \$50	Mostly deceased	Severe male mortality
Solo & Small	3rd	0	< \$50	Mixed	High numbers in both of survivor/deceased categories
Small Family	1 st , 2 nd	1, 0	\$50 - \$100, < \$50	Mostly survived	High female survivor rate
Solo	1 st	0	< \$50	Mixed	High numbers in both of survivor/deceased categories

Title, Family Group, Cabin Count, Fare, Survived

Title



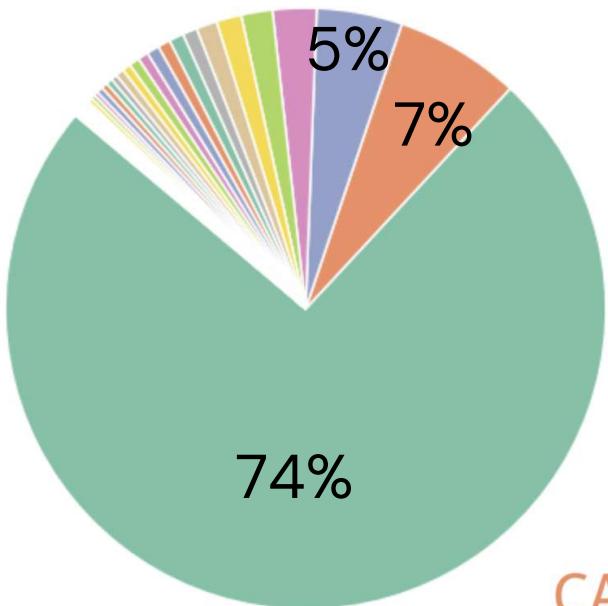
Title



Ticket Group

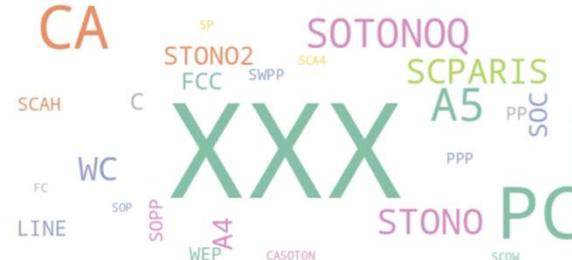
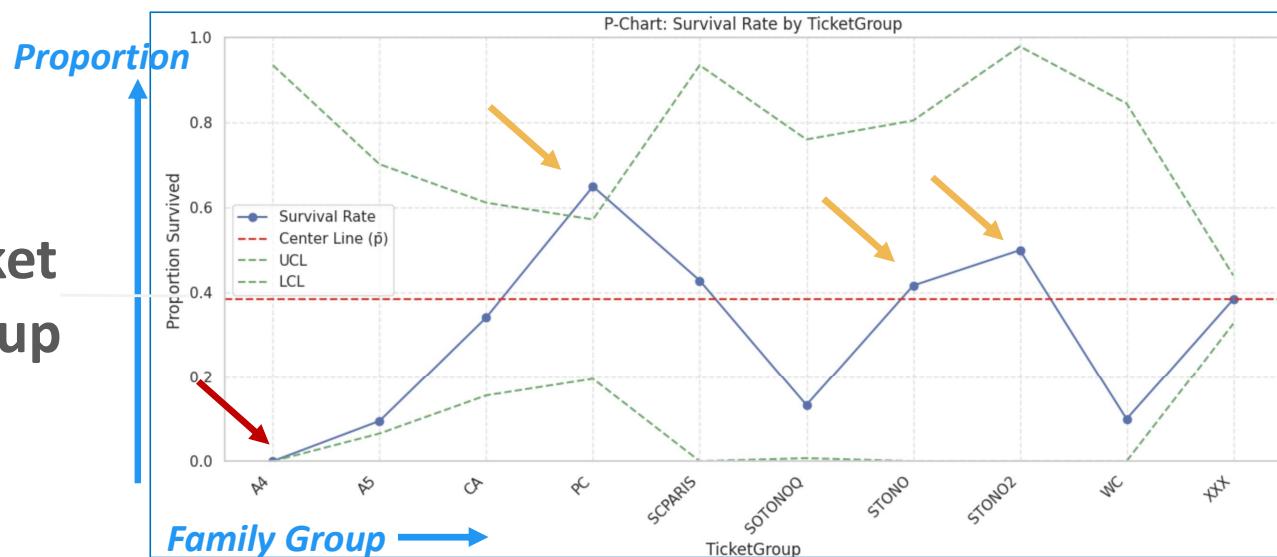
Key takeaways:

Most passenger's ticket only display numbers, followed by those with PC and CA



Insights

Ticket Group



Key takeaways:
Ticket with 'PC' had higher survival rate

Titanic Data



Continuous



Age



AgeGroup



Fare_log



FareGroup



Baseline – Features Selection

Categorical



Title



PClass



Surname



CabinGroup



Sex



CabinMGroup



Embarked



TicketGroup



Count



Parch



SibSp



CabinCnt



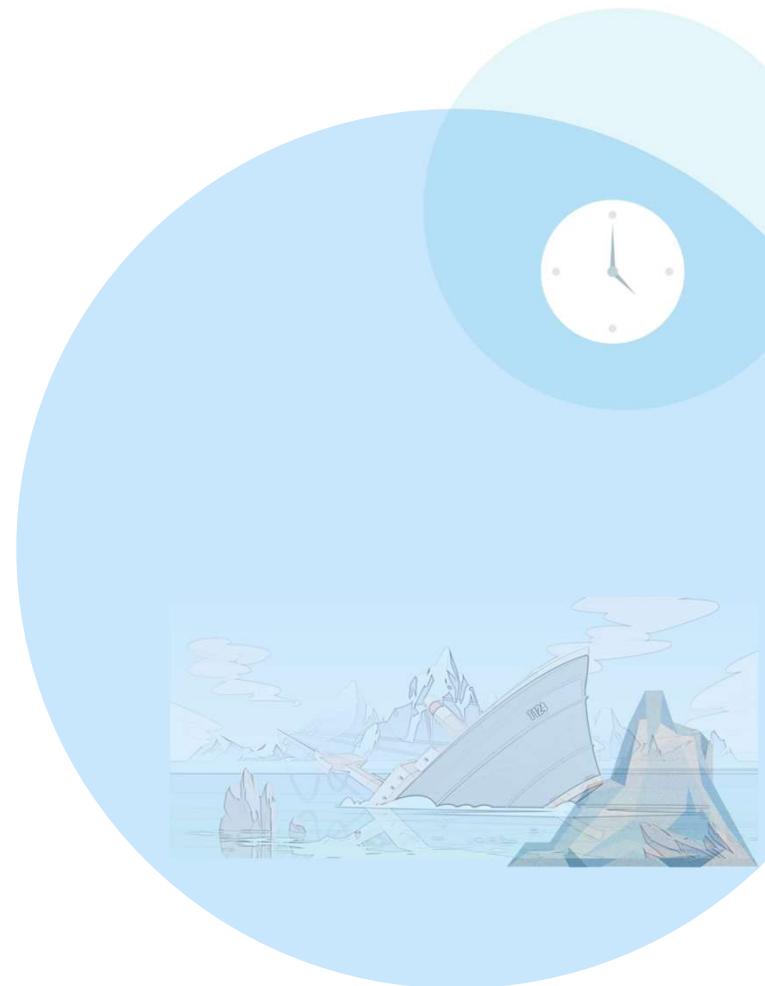
FamilySize

13 Features → 83 Features

5

Construct

Construct and Evaluate the Model



Titanic Data

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	train_set	Survived
1171	2	Oxenham, Mr. Percy Thomas	male	22.0	0	0	W./C. 14260	10.5000	NaN	S	0	NaN
962	3	Mulvihill, Miss. Bertha E	female	24.0	0	0	382653	7.7500	NaN	Q	0	NaN
976	2	Lamb, Mr. John Joseph	male	NaN	0	0	240261	10.7083	NaN	Q	0	NaN
928	3	Roth, Miss. Sarah A	female	NaN	0	0	342712	8.0500	NaN	S	0	NaN
980	3	O'Donoghue, Ms. Bridget	female	NaN	0	0	364856	7.7500	NaN	Q	0	NaN
1306	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9000	C105	C	0	NaN
1021	3	Petersen, Mr. Marius	male	24.0	0	0	342441	8.0500	NaN	S	0	NaN
939	3	Shaughnessy, Mr. Patrick	male	NaN	0	0	370374	7.7500	NaN	Q	0	NaN
894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q	0	NaN
1233	3	Lundstrom, Mr. Thure Edvin	male	32.0	0	0	350403	7.5792	NaN	S	0	NaN

all_data_df

train_df (test.csv)

test_df (test.csv)

10 Features (X)

83 Features (X)

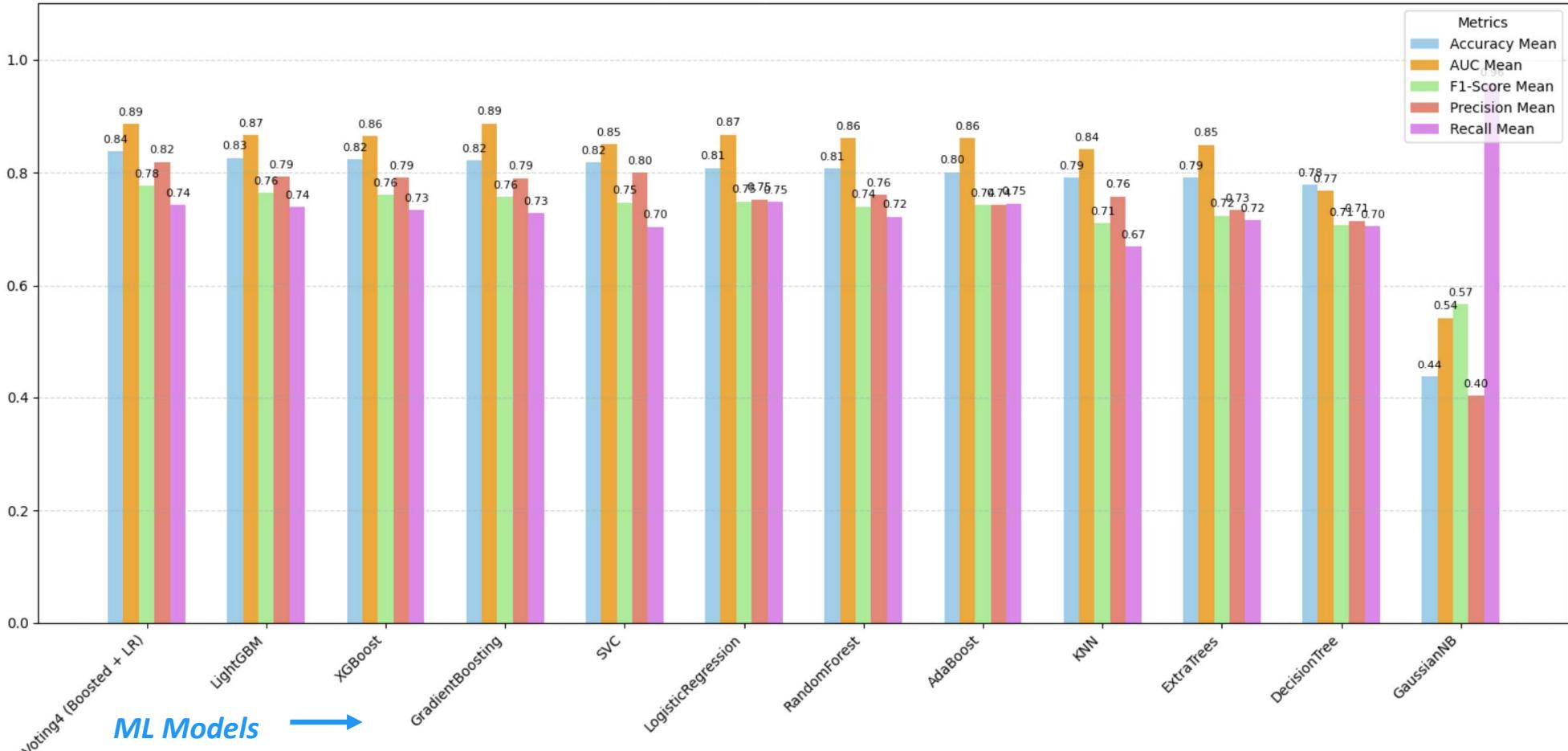
Target

Y_train

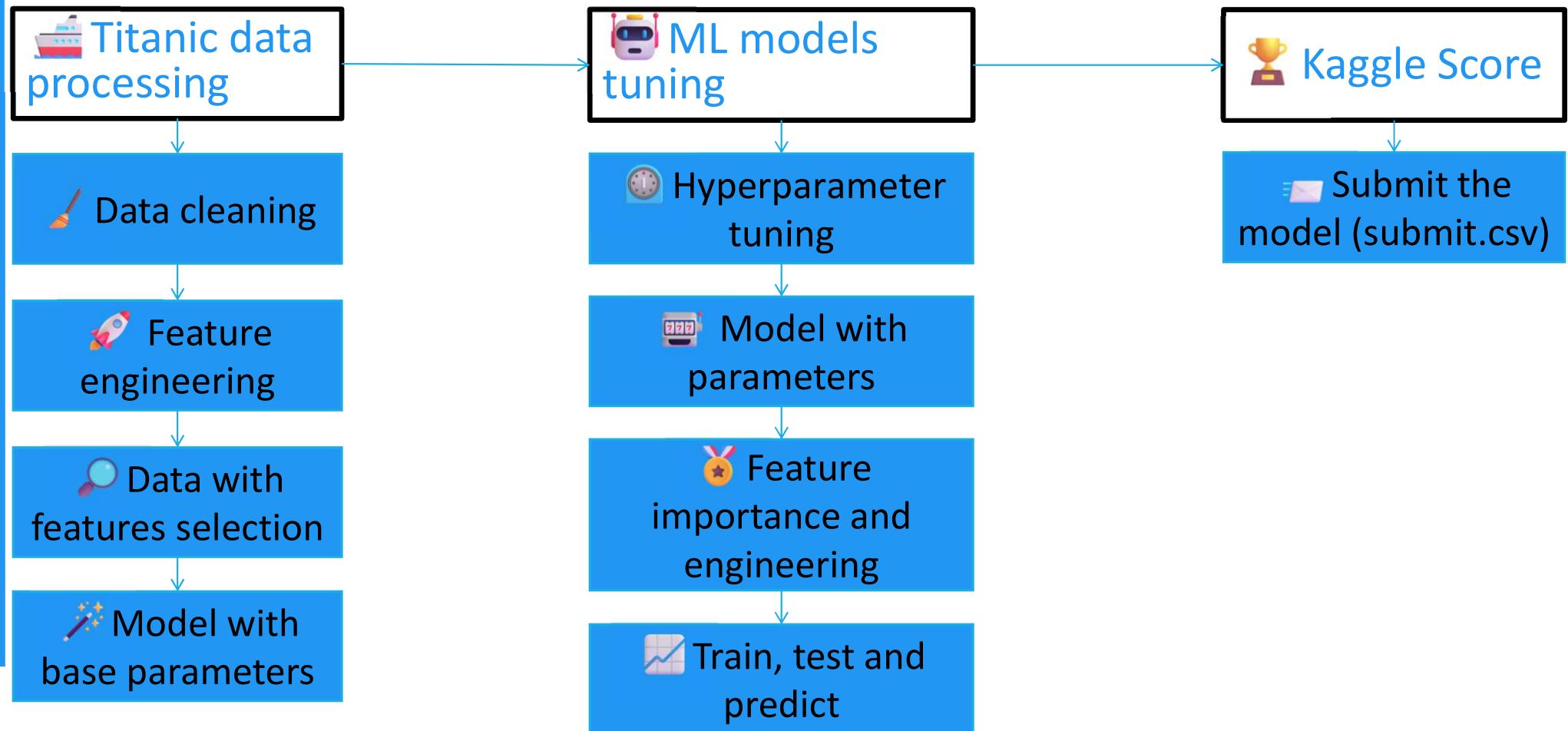
Titanic Data

Mean Score

12 Models
Model Comparison by Evaluation Metrics (Mean)

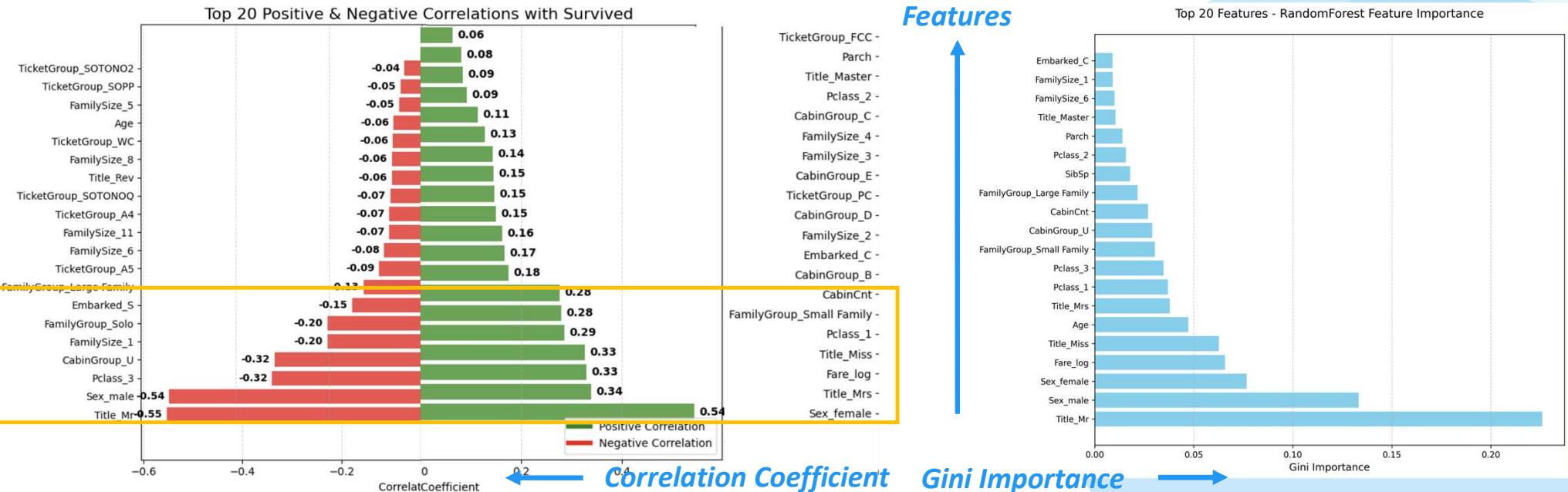


Flow Chart



Titanic Data

Baseline – Features Selection



 Title

 Fare_log

 Sex

 PClass

17  CabinGroup

 CabinCnt

 Embarked

 FamilyGroup

 Age

Titanic Data

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	train_set	Survived
1171	2	Oxenham, Mr. Percy Thomas	male	22.0	0	0	W./C. 14260	10.5000	NaN	S	0	NaN
962	3	Mulvihill, Miss. Bertha E	female	24.0	0	0	382653	7.7500	NaN	Q	0	NaN
976	2	Lamb, Mr. John Joseph	male	NaN	0	0	240261	10.7083	NaN	Q	0	NaN
928	3	Roth, Miss. Sarah A	female	NaN	0	0	342712	8.0500	NaN	S	0	NaN
980	3	O'Donoghue, Ms. Bridget	female	NaN	0	0	364856	7.7500	NaN	Q	0	NaN
1306	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9000	C105	C	0	NaN
1021	3	Petersen, Mr. Marius	male	24.0	0	0	342441	8.0500	NaN	S	0	NaN
939	3	Shaughnessy, Mr. Patrick	male	NaN	0	0	370374	7.7500	NaN	Q	0	NaN
894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q	0	NaN
1233	3	Lundstrom, Mr. Thure Edvin	male	32.0	0	0	350403	7.5792	NaN	S	0	NaN

all_data_df

train_df (test.csv)

test_df (test.csv)

10 Features (X)

83 Features (X)

17 Features (X)

X_train_reduced

X_test_reduced

Target

Y_train

Titanic Data

Random Forest and SVC

Model	Description	Cross-validation score	Kaggle score
Random Forest(1)	83 features, lightly imbalanced data, X_train	83.7%	78.5% 
Random Forest(2)	83 features, balanced data , X_train	82.9% 	76.8% 
SVC(1)	83 features, lightly imbalanced data, X_train, rbf	82.7% 	76.1% 
Random Forest(3)	17 features, lightly imbalanced data, X_train_reduced_1	83.5% 	79.2% 
Random Forest(4)	83 features, lightly imbalanced data, X_train_1, use 'Fare'	83.7% 	78.5% 
Random Forest(5)	72 features, lightly imbalanced data, X_train_2, 'Title_Less'	83.7% 	78.5% 
Random Forest(6)	17 features, lightly imbalanced data, X_train_2 X_train_reduced_2, 'Title_Less'	83.5% 	79.2% 

Titanic ML Models



Random Forest

Model	Description	Cross-validation score	Kaggle score
Random Forest(7)	86 features, lightly imbalanced data, X_train_3, 'Survived_predict'	83.2%	78.9%
Random Forest(8)	19 features, lightly imbalanced data, X_train_reduced_3, 'Survived_predict'	83.2%	77.8%
Random Forest(9)	10 features (base), lightly imbalanced data, X_train_reduced_4, threshold =0.03	83.2%	77.5%
Random Forest(10)	13 features (base), lightly imbalanced data, X_train_reduced_5, threshold =Pareto	83.5%	77.3%
Random Forest(11)	21 features (base), lightly imbalanced data, X_train_reduced_6, threshold =0.075	83.5%	77.9%
Random Forest(28)	19 features, lightly imbalanced data, X_train_reduced_3, 'Survived_predict', max_depth = 10	83.2%	79.7%

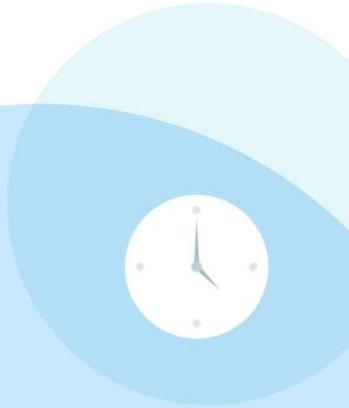
Titanic Data

Ensemble and Logistic Regresion

Model	Description	Cross-validation score	Kaggle score
Logistic regression(1)	83 features, lightly imbalanced data, X_train_1	81.9%	77.3%
Logistic regression(2)	17 features, lightly imbalanced data, X_train_r_1	82.5%	78.5% 
XG boost (1)	83 features, lightly imbalanced data, X_train_1	82.4%	76.8%
XG boost (2)	17 features, lightly imbalanced data, X_train_r_1	82.8%	76.3%
Voting classifier (boosting + LR)	83 features, lightly imbalanced data, X_train_1	82.5%	77.8%
Gradient boosting (1)	83 features, lightly imbalanced data, X_train_1	83.3%	78.9% 
Gradient boosting (2)	17 features, lightly imbalanced data, X_train_r_1	82.8%	78.2%
KNN	17 features, lightly imbalanced data, X_train_r_1	80.4%	74.6%
Ada boosting (1)	83 features, lightly imbalanced data, X_train_1	79.7%	75.8%

6

Execute Conclusion



OKR Status



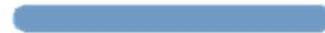
Objective

Develop a High-Performing Survival Prediction Model

Status

In Progress

Progress Score



Deadline

May 1, 2025

Key Result 1

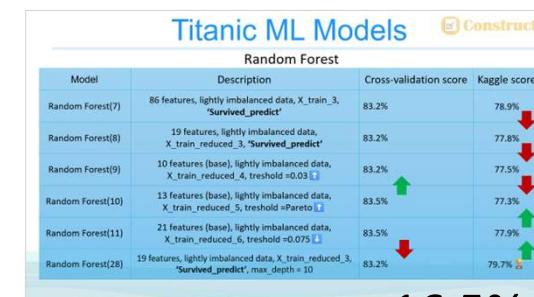
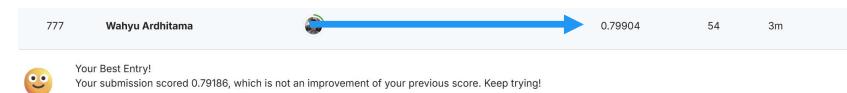
Achieve Kaggle submission score >75%

Key Result 2

Reach validation accuracy, precision, recall, and F1-score >82%

Key Result 3

Final model has generalization gap <5% between training and Kaggle score



16.5%

OKR Status



Benchmark and Evaluate
Multiple ML Models

Objective

Next

Status

Progress Score

May 5, 2025

Deadline

Key Result 1

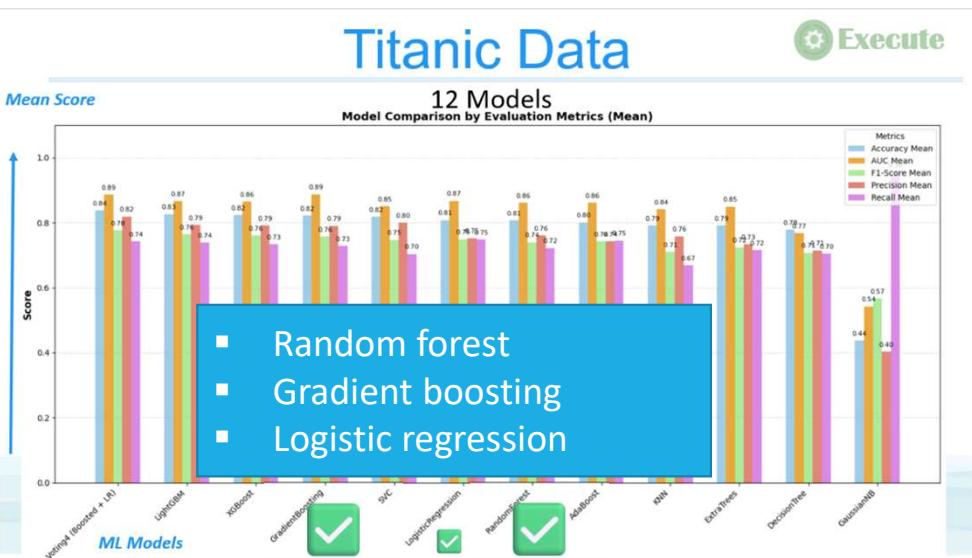
Train the promising machine
learning models

Key Result 2

Identify top 3 models based on
accuracy and precision

Key Result 3

Reduce runtime for model
training by 20%



OKR Status

Objective

Status

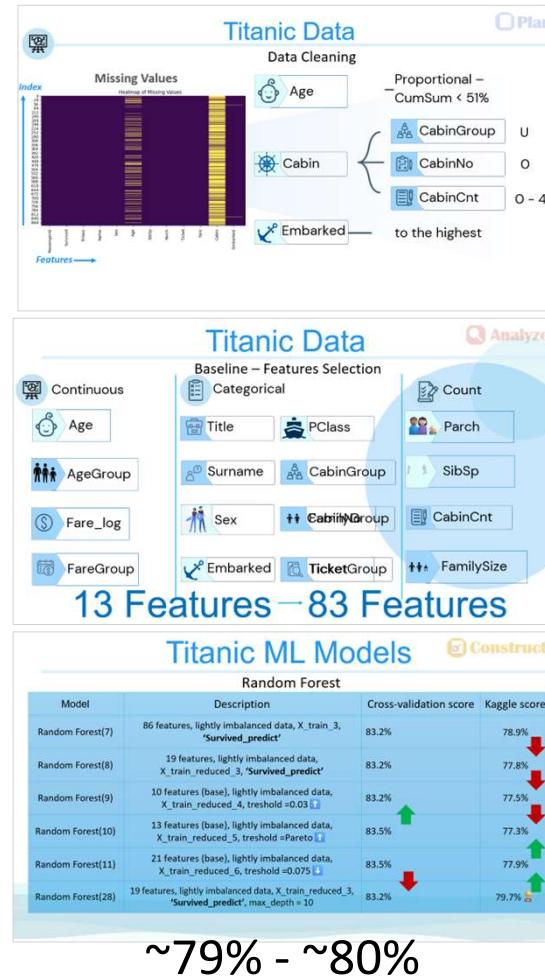
Progress Score

Deadline

Key Result 1

Key Result 2

Key Result 3



Enhance Feature Engineering and Data Quality

Next

May 7, 2025

Reduce missing data to <1% across all relevant features

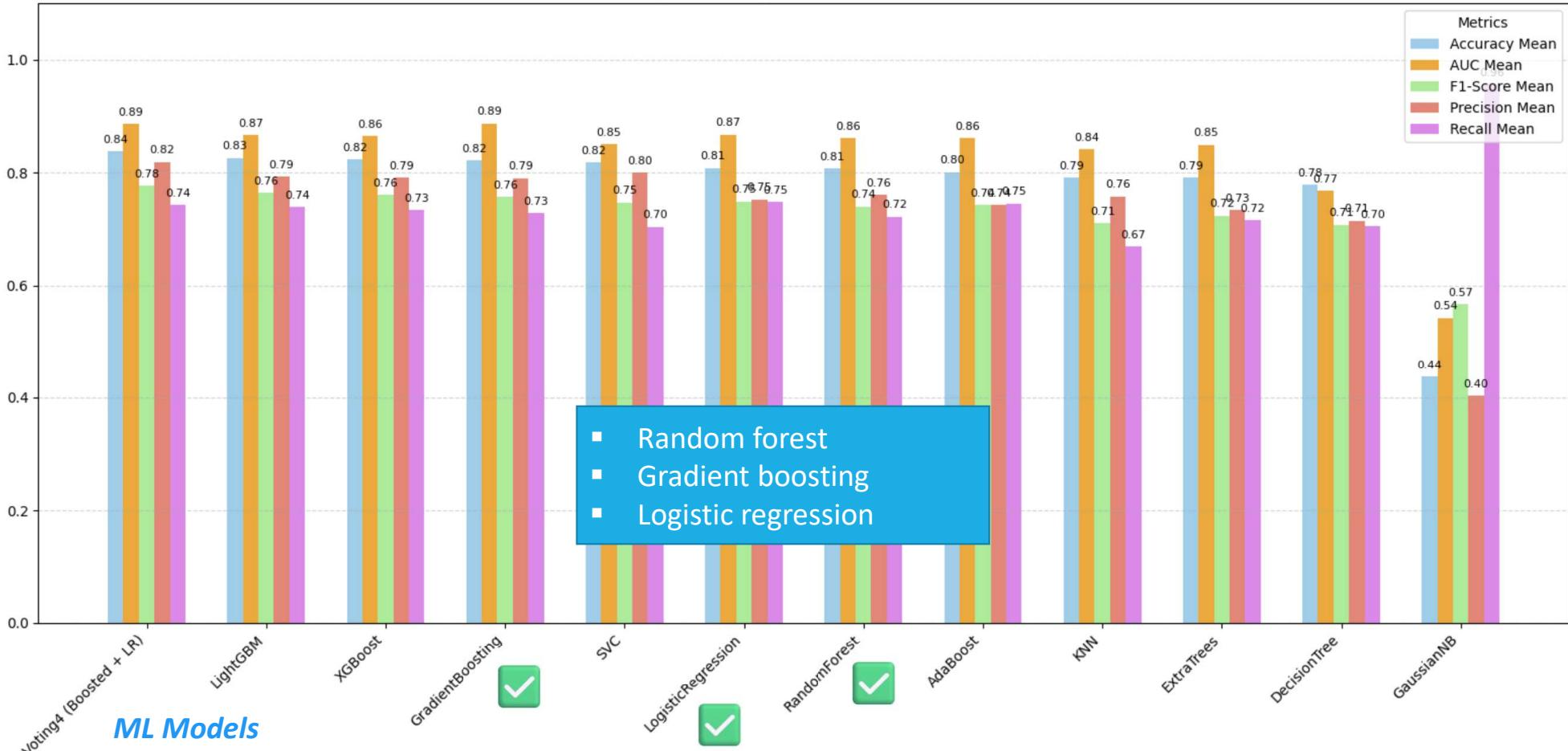
Create at least 5 new meaningful engineered features

Improve model performance by at least 5% after feature engineering

Titanic Data

Mean Score

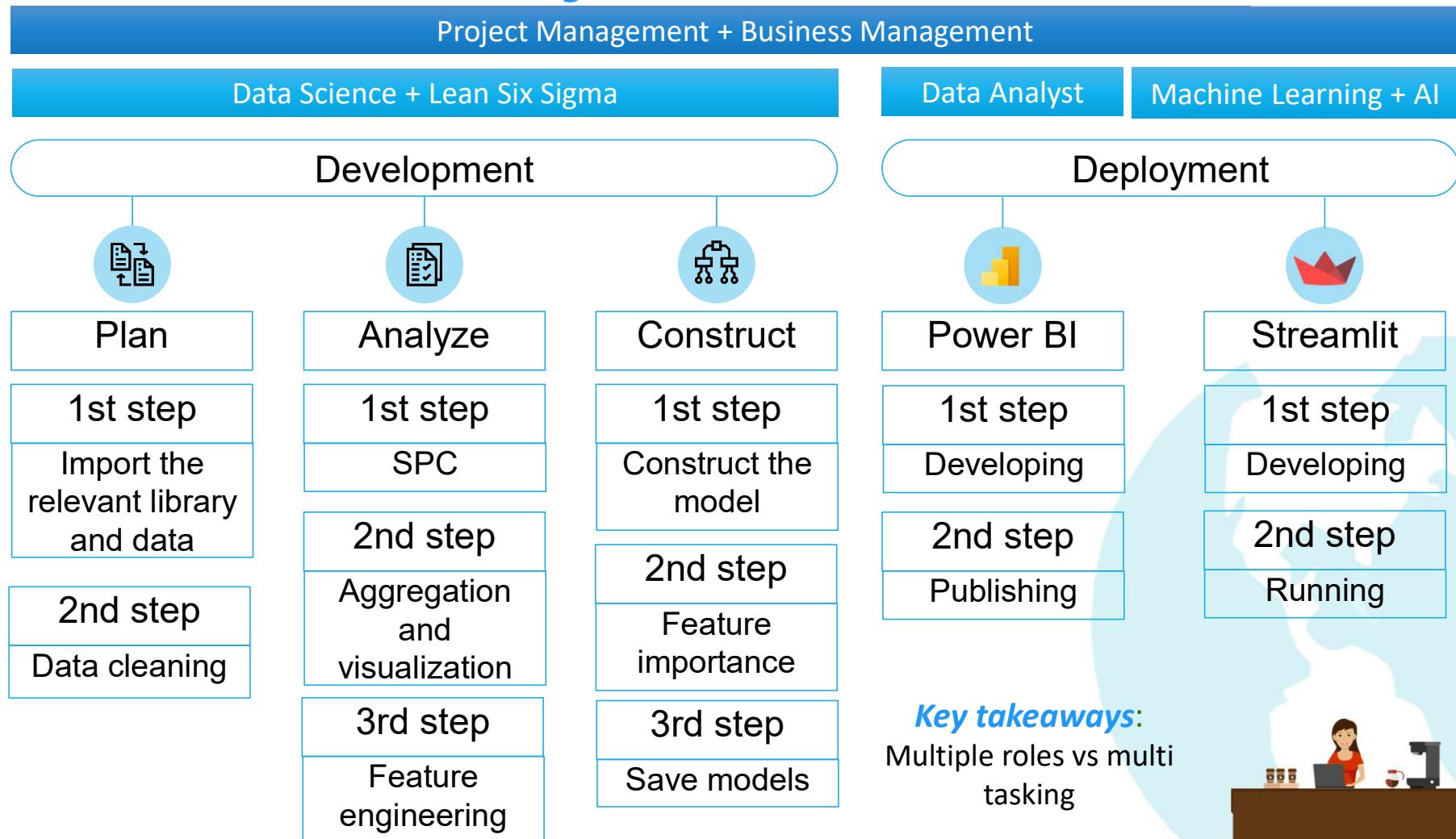
12 Models
Model Comparison by Evaluation Metrics (Mean)



Data Project Architecture



Execute



Next Steps

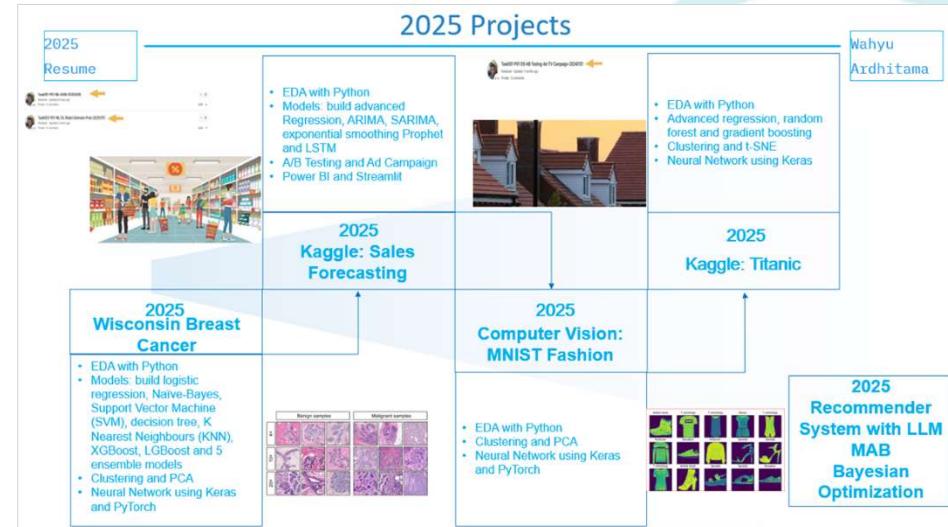


Improvements:

- Expand model testing beyond random forest to include:
 - Gradient boosting: capture complex non-linear relationships
 - Logistic regression: provides interpretable results and serves as baseline
- Enhanced feature engineering, Fare_group categories to capture price ranges and their correlation with survival
- Voting classifier combining random forest, gradient boosting, and logistic regression

Next projects:

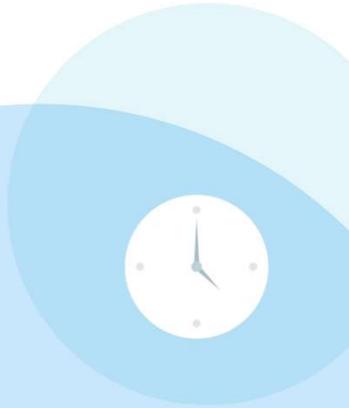
- Wisconsin breast cancer
- Kaggle: sales forecasting
- Computer vision: MNIST – fashion
- Kaggle: house price



i

Appendix

Videos, Links and Documentations



Appendix

Documents:

- <https://www.kaggle.com/code/wahyuardhitama/task001-p01-ml-titanic-20250407>
- <https://www.kaggle.com/code/wahyuardhitama/task001-p02-ml-titanic-20250416>
- <https://www.kaggle.com/code/wahyuardhitama/task001-p03-ml-titanic-20250506>
- <https://www.kaggle.com/code/wahyuardhitama/task001-p04-ml-titanic-20250507>
- <https://github.com/whyzie/Task001-ML-Titanic-20250407>

