TITANIC

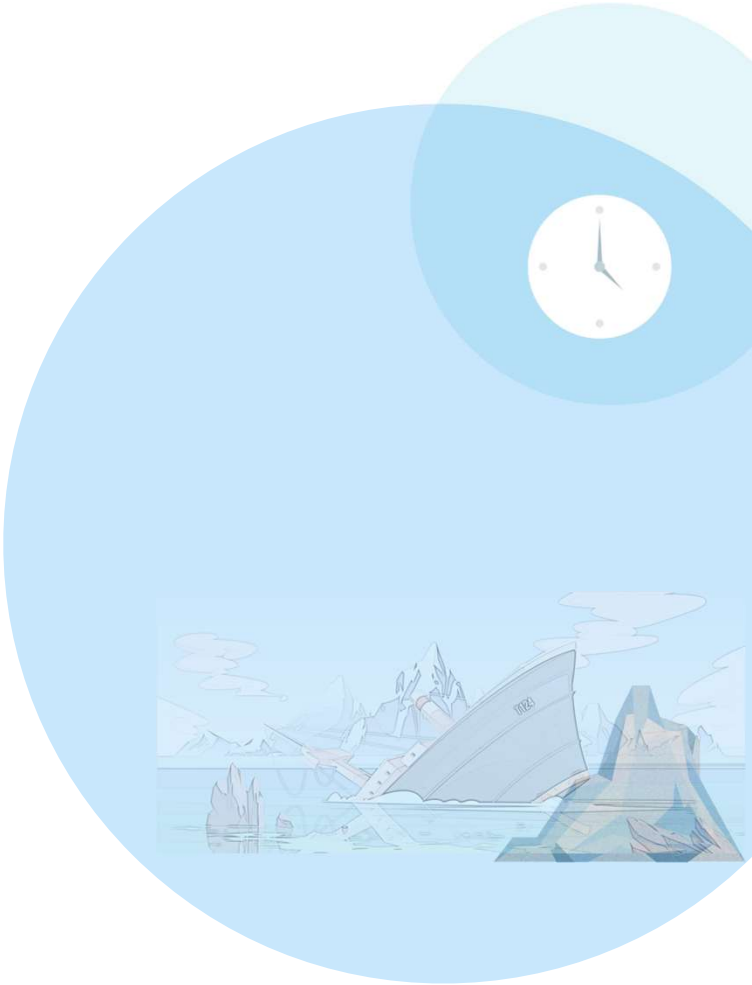
Presented by: Wahyu Ardhitama
Last Updated: May 13th, 2025
#Task001-P01-ML-Titanic-20250407

Table of Contents

1

Introduction

Data Science and Machine Learning Resume



Data Project Architecture

Project Management + Business Management

Data Science + Lean Six Sigma

Data Analyst

Machine Learning + AI

Development

Plan

1st step

Import the relevant library and data

2nd step

Data cleaning

Analyze

1st step

SPC

2nd step

Aggregation and visualization

3rd step

Feature engineering

Construct

1st step

Construct the model

2nd step

Feature importance

3rd step

Save models

Deployment

Power BI

1st step

Developing

2nd step

Publishing

Streamlit

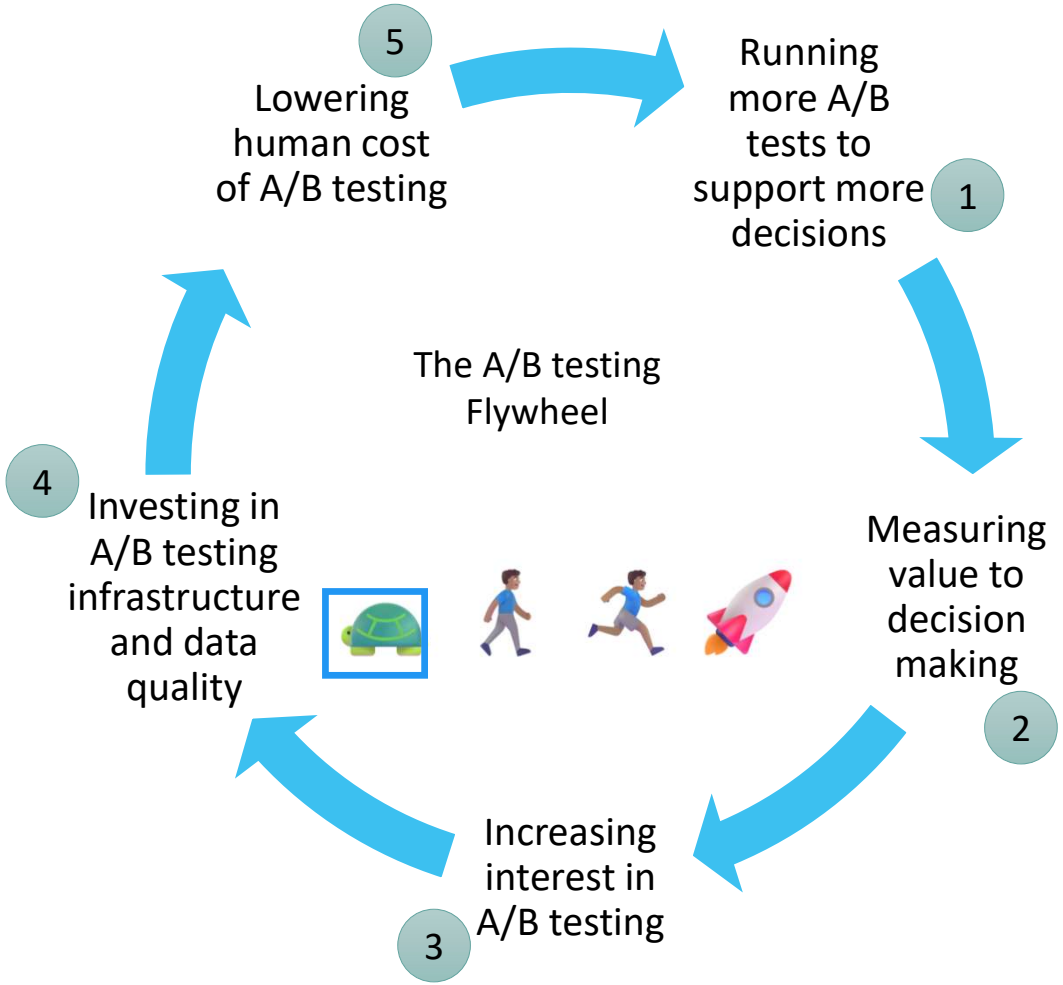
1st step

Developing

2nd step

Running

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

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

Project Planning

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

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

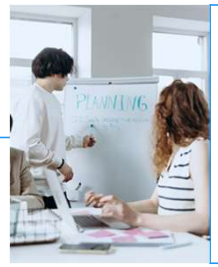
Project Online Meetings

01

Gather Consenssus

Will need your feedback – Kick Off

Stakeholder requirements
Project requirements
Strategy documents
Mock up



Project Initiation

Prepare- discover our end goal

02

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

Project Execution

Start to race

04

Customer Survey
ROAM Analysis
Status Report
Project Closure

Wahyu
Ardhitama

Analysis and Experimentation Team

Mission:

1. Develop a platform that is easy to integrate across systems
2. **F**oster a data-driven culture by enabling fact-based and analytical decision-making
3. **A**ccelerate innovation through reliable experimentation and insightful result analysis
4. **M**aximize ROI by aligning tasks and projects with business objectives
5. **E**mpower 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: **C**lear/**C**lean/**C**ommunication/**C**ollaboration/**C**orrection, **O**bjective, **V**aluable, **F**ocus, **A**gile, **S**cientific and **T**ime-bound/Trustworthiness

2025 Resume

Projects

Wahyu Ardhitama

- 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

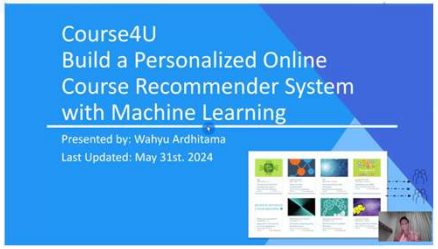
2025
Kaggle: Titanic

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



2025
Resume

Skills

Wahyu
Ardhitama

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

●●○○○ Deep Learning Models

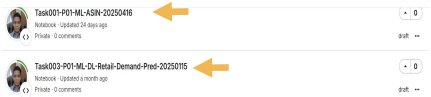
04

LSTM, NLP
CNN, Computer Vision
LLM, RAG ●○○○○

2025 Projects

2025 Resume

Wahyu Ardhitama



- 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

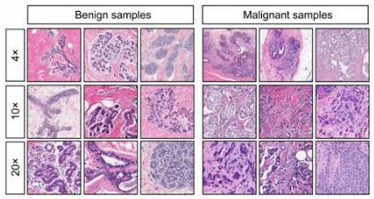
- EDA with Python
- Advanced regression, random forest and gradient boosting
- Clustering and t-SNE
- Neural Network using Keras

2025 Kaggle: Titanic

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

+ Chest Cancer



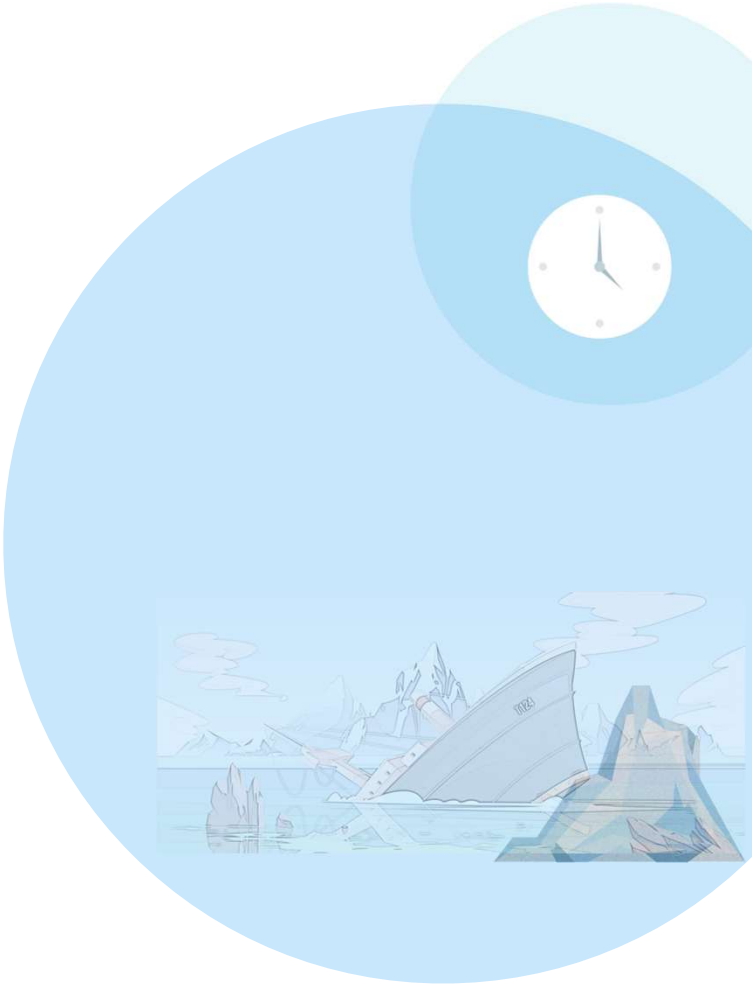
2025 Computer Vision: MNIST Fashion

- 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...

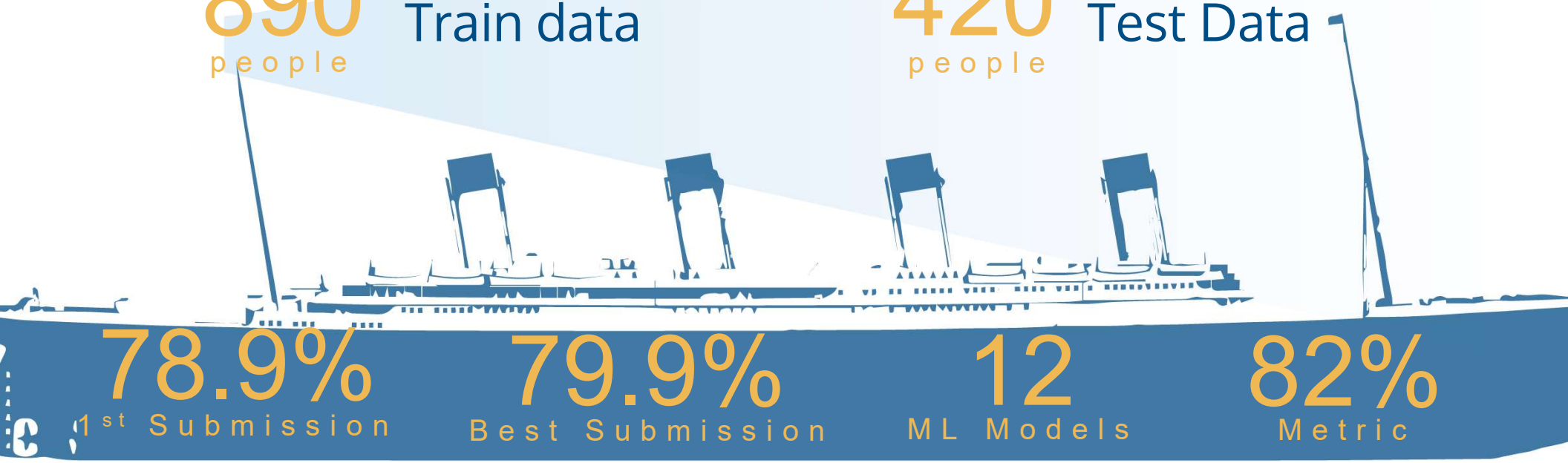
0 Submissions left today

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

Plan

Predicting Titanic survivors based on passenger data

Analytics Objective

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

Titanic Survival Prediction

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

Hypothesis

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

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	❑ PACE framework ❑ Agile project management ❑ Lean 7 Waste + SPC + Six Sigma
Goals & Improvement metric	○ **Primary : More than 79% score** ○ **Guardrail : 82% accuracy; 80% precision**

Titanic Competition

SIPOC

S Supplier	I Input	P Process	O Output	C Customer
1. Data Analyst 2. Safety Board 3. Kaggle Competition	1. Raw Data 2. OKR	Plan 1. Data Cleaning 2. EDA 3. Feature Engineering 4. Construct Model 5. Presentation 6. Submit	1. Pre-process data 2. Insights 3. ML models 4. Kaggle score	1. Safety Board 2. Kaggle

Process Waste

Plan

The 7 Wastes Plus

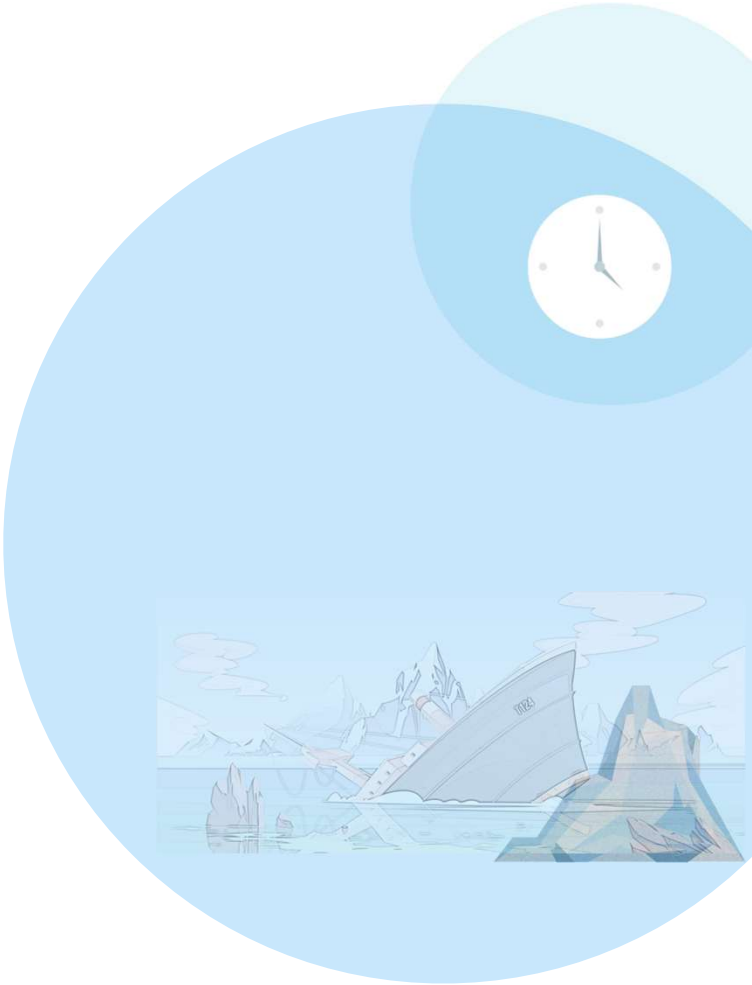
7 Waste + In Lean (Muda)	Findings	Mitigate Strategy
Waiting	Long training time for complex models (e.g., LGBM, XGBoost, Gradient Boosting, Voting Classifier)	Use faster models (e.g., Random Forest), subset data, or use GPU acceleration and parallel processing
Inventory	Storing too many intermediate models, 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			
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% ✅	▪ 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

Development – 1 to 2 weeks

Deployment 1 – 2 weeks

Plan	Analyze	Construct	Power BI	Kaggle	Streamlit
1st step Import the relevant library and data	1st step SPC	1st step Construct the model	1st step Developing	1st step Submitting	1st step Developing
2nd step Data cleaning	2nd step Aggregation and visualization	2nd step Feature importance	2nd step Publishing		2nd step Running
	3rd step Feature engineering	3rd step Save models			

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

Current

Continuous

- Age
- Fare

Categorical

- Name
- PClass
- Sex
- Cabin
- Embarked
- Ticket

Count

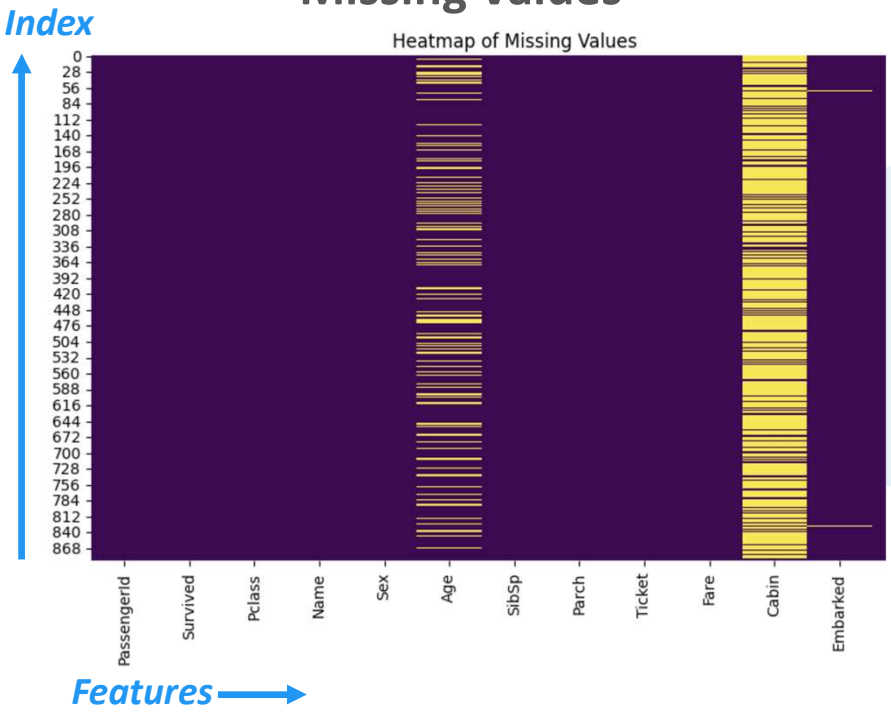
- Parch
- SibSp

Survived

Titanic Data

Data Cleaning

Missing Values



- **Age** – Proportional – CumSum < 51%
- **Cabin**
  - CabinGroup — U
  - CabinNo — 0
  - CabinCnt — 0 - 4
- **Embarked** — to the highest

Titanic Data

Data Wrangling

Continuous

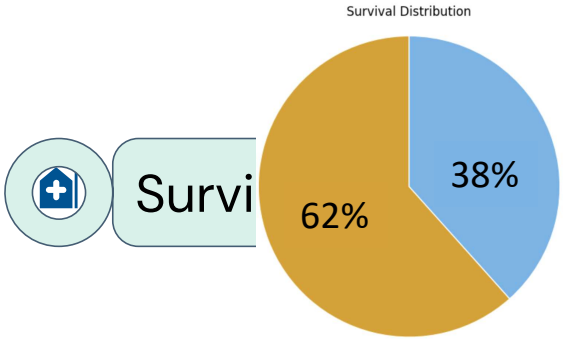
- Age
- Fare

Categorical

- Name
- PClass
- Sex
- CabinGroup
- Embarked
- TicketNo

Count

- Parch
- CabinCnt



Key takeaways

Survived passengers = 340 (38%)

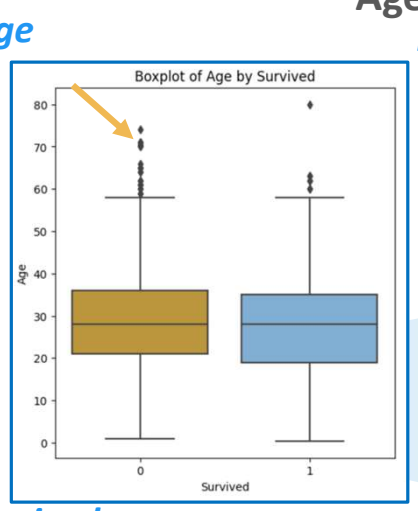
Deceased passengers = 550 (62%)

Age and Fare

EDA

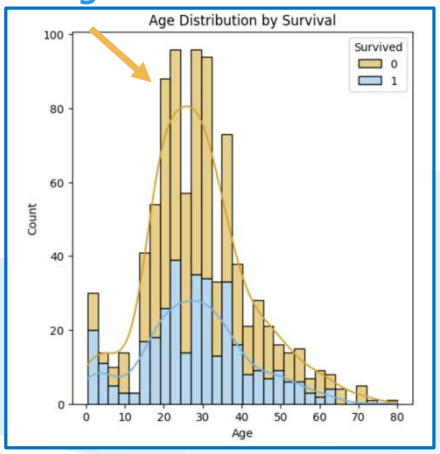
Age

Age



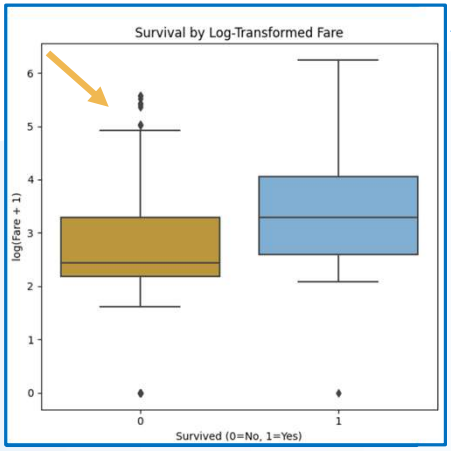
Survived

Passengers



Age

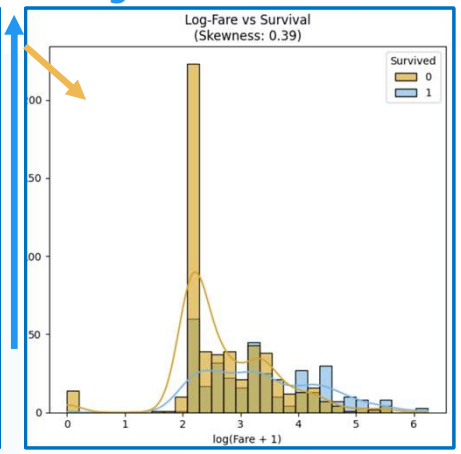
Fare log



Survived

Log-Fare

Passengers



Fare log

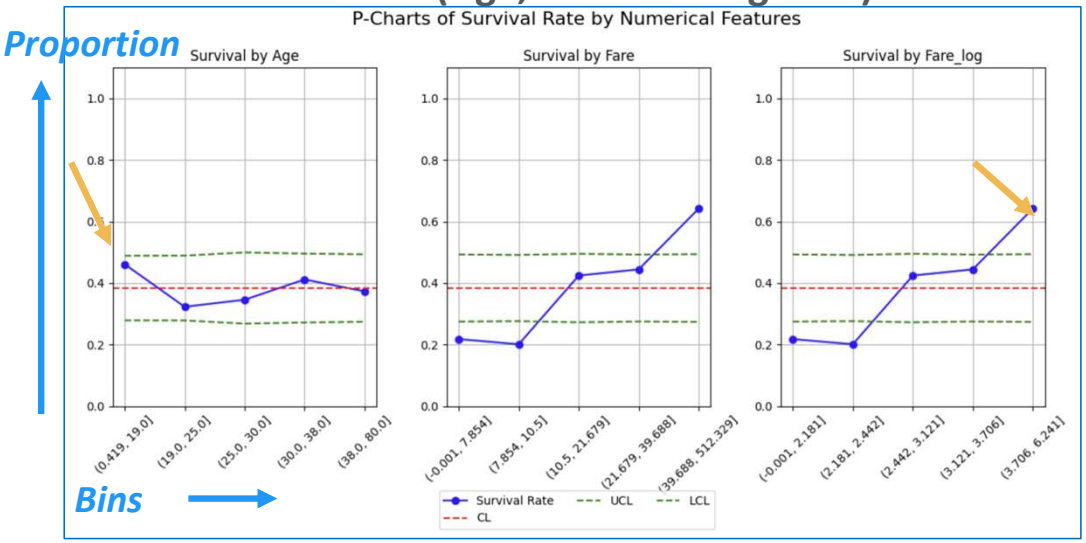
Practical Interpretations:

	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

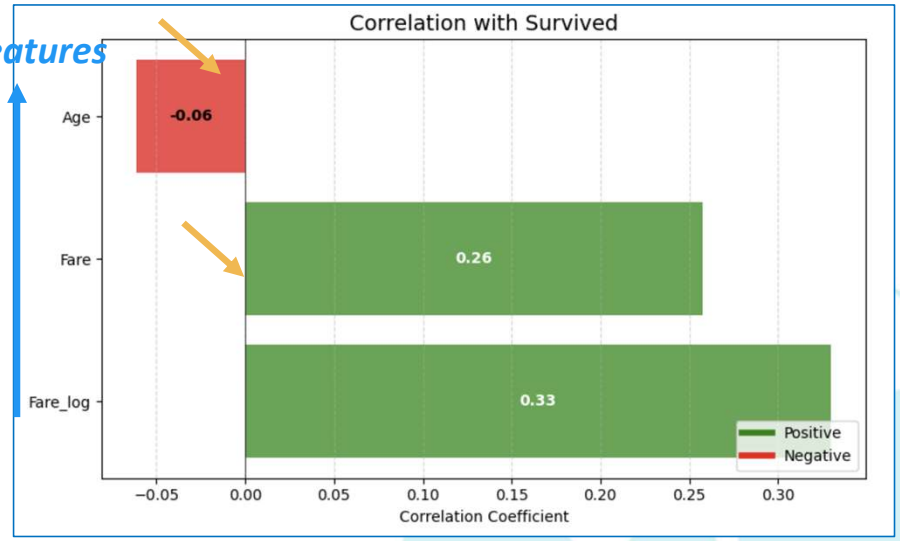
Age and Log-Fare

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%)

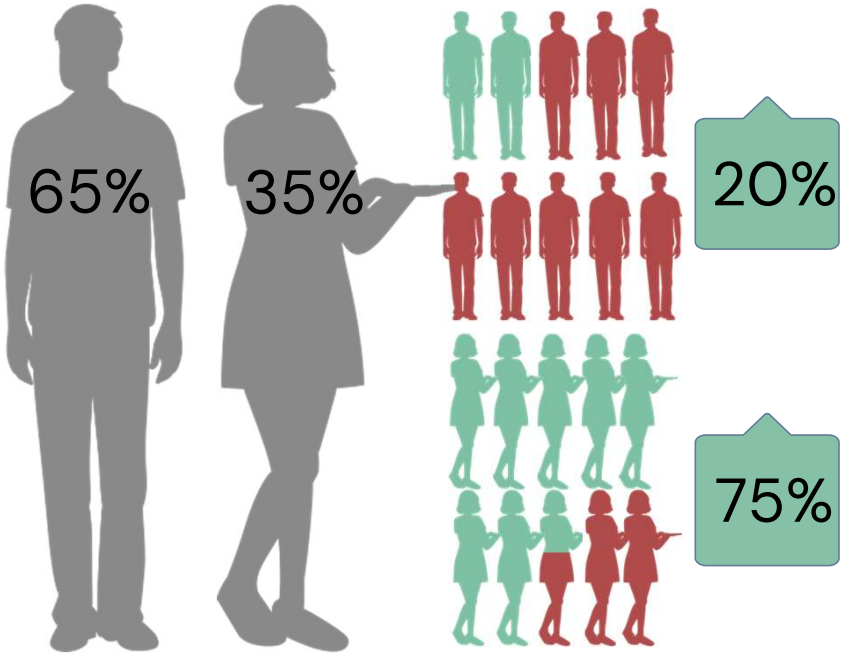
Next: Elaborate 'Age' into child (under 13) adolescent (13 – 19), adult (19-64) and seniors

Sex and Pclass

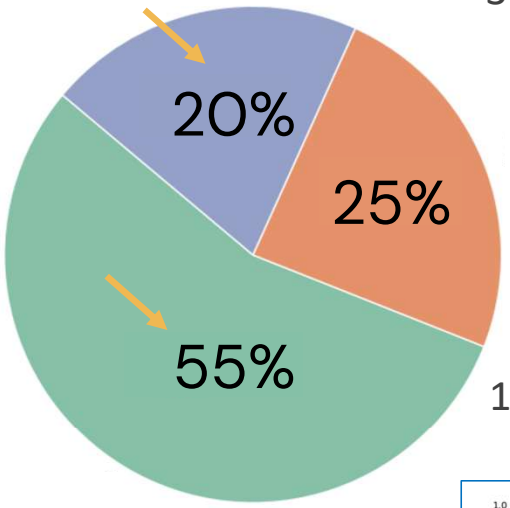
Insights

Key takeaways:

A higher proportion of females survived



Sex



Key takeaways:

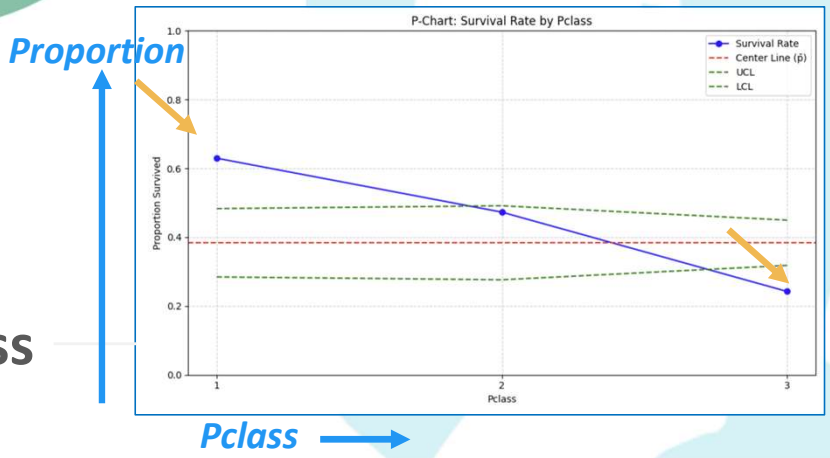
3rd Class had the highest passengers

2nd class had the lowest

Key takeaways:

1st Class had better survival chance

3rd class had the lowest



Proportion

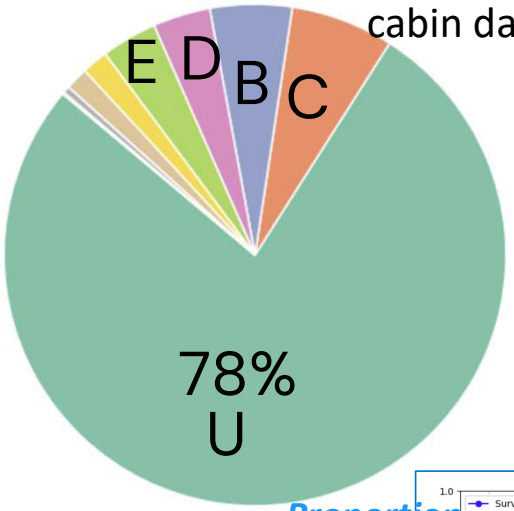
Pclass

Pclass

Cabin Group and Cabin No

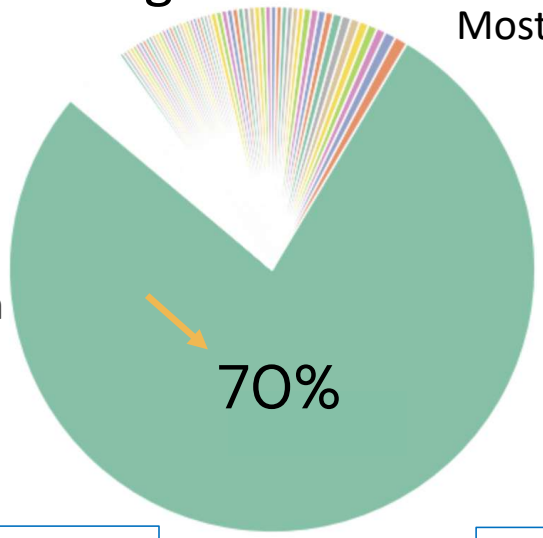
Analyze

Key takeaways:
Most passengers fall into the group with unknown cabin data



Key takeaways:
Passengers from cabin A to G had a better survival chance.
(The top B, D and E)

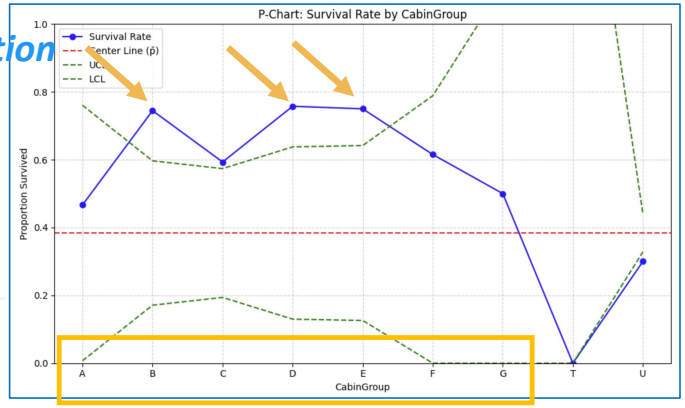
Insights



Key takeaways:
Most passengers Cabin's Number is 0, unknown Cabin

Key takeaways:
Cabin's Numbers come in small group with highest total is 7
Number 0 has below 40% survival chance

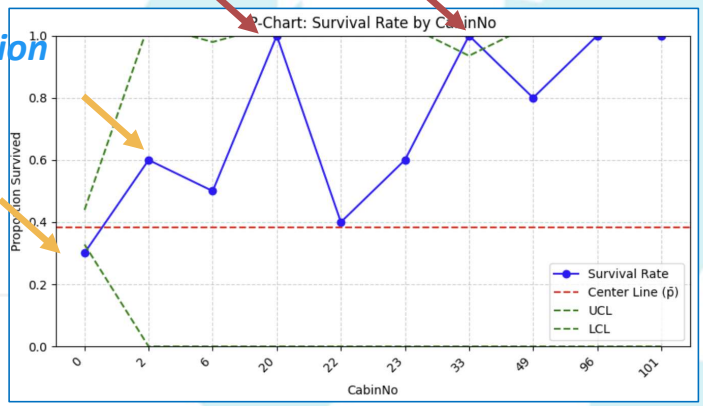
Proportion



Cabin Group

Embarked →

Proportion



Cabin No

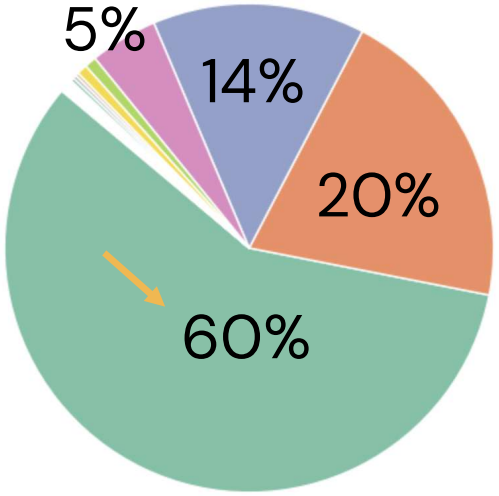
Embarked →

Title

Analyze

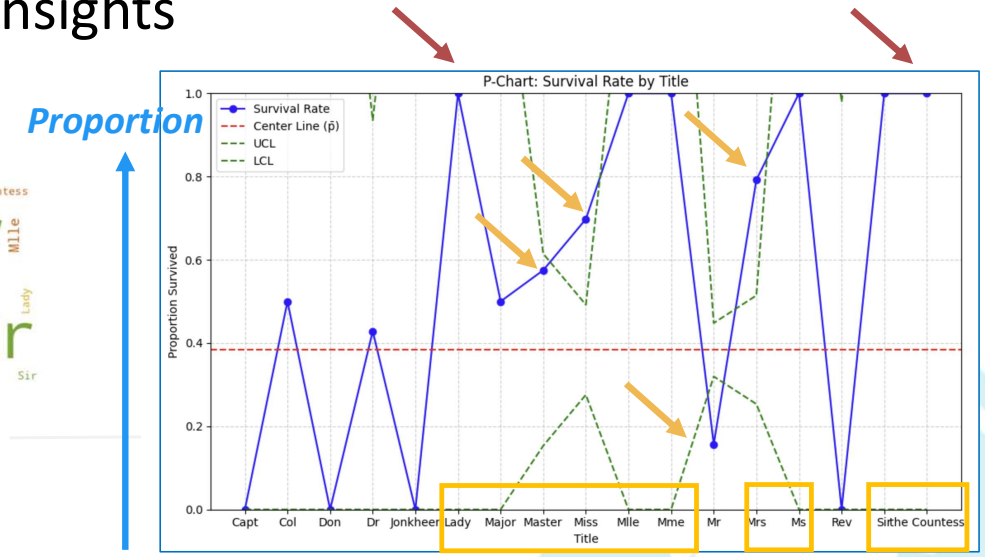
Key takeaways:

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



Title

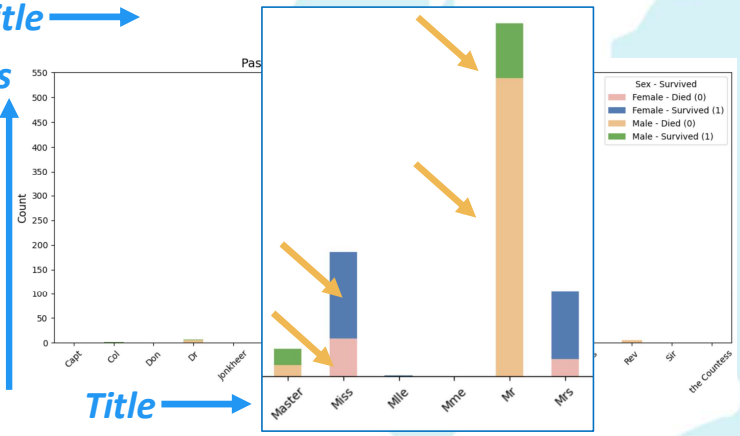
Insights



Proportion

Title

Passengers



Key takeaways:

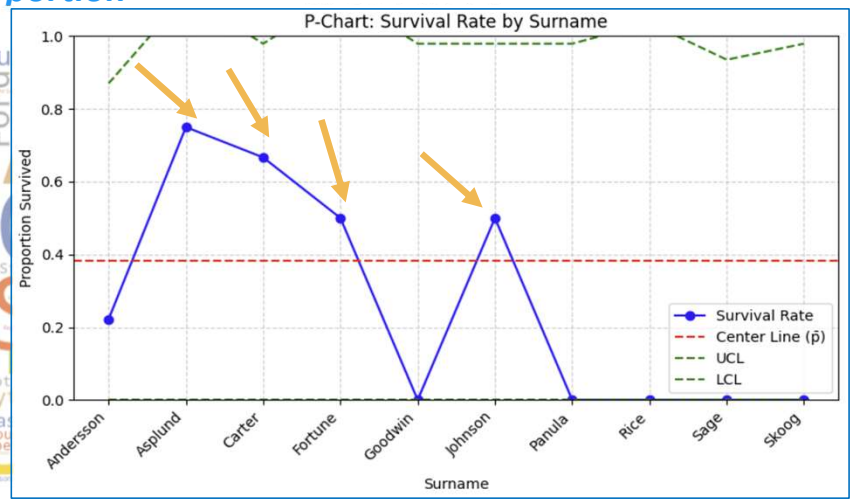
Female passengers with titles such as Miss (70%), Mrs (80%), and noble titles had higher survival rates, while ‘Mr’ (15%) and ‘Master' ’(55%) were among the most common male titles.

Surname and Ticket

Insights

Surname

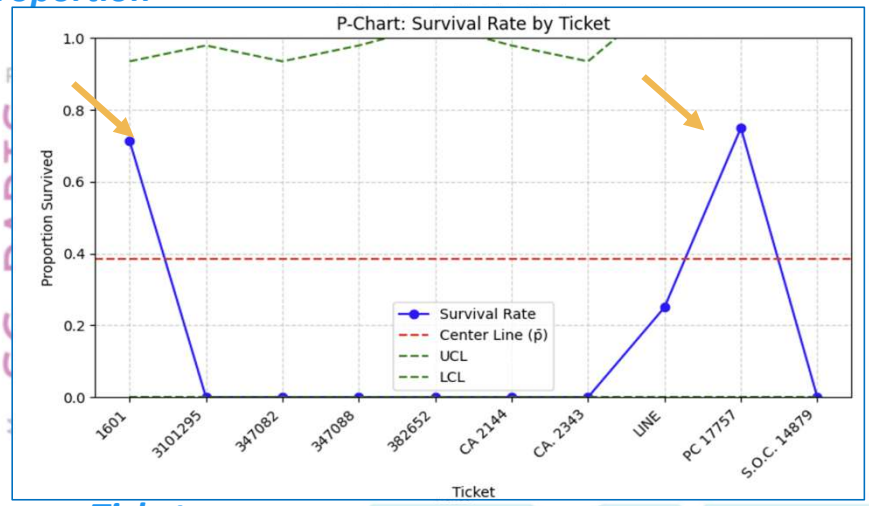
Proportion



Surname →

Ticket

Proportion



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

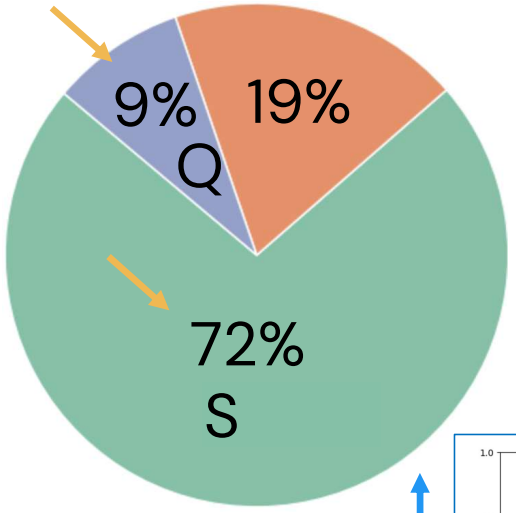
Passengers with surnames such as **Asplund, Carter, Fortune**, and **Johnson** appear to have had **better chances of survival**. Anderson appearing in up to 9 members with only 2 survived.

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

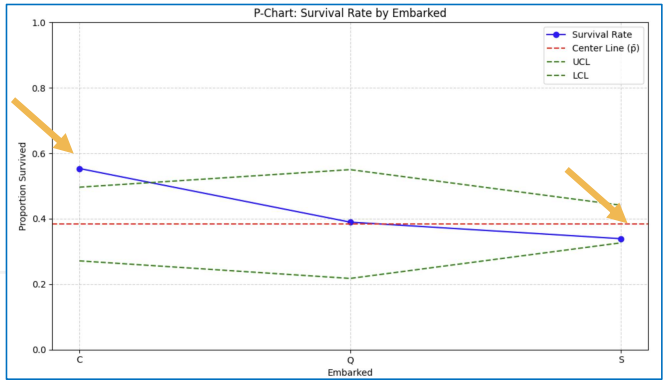
Embarked and Multiple Cabin

Insights

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



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

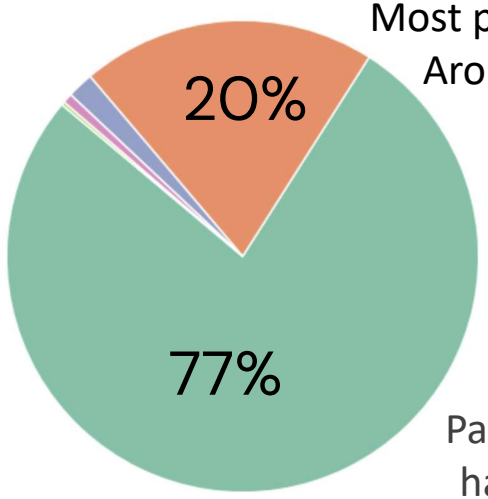


Proportion

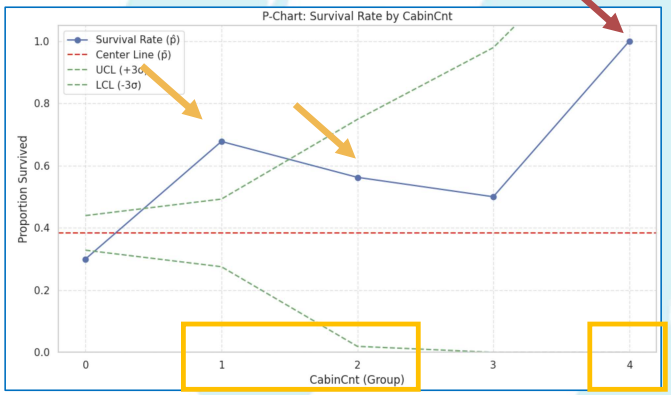
Embarked

Embarked →

Key takeaways:
Most passengers had no cabin
Around 20% had 1 cabin



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



Proportion

Cabin Count

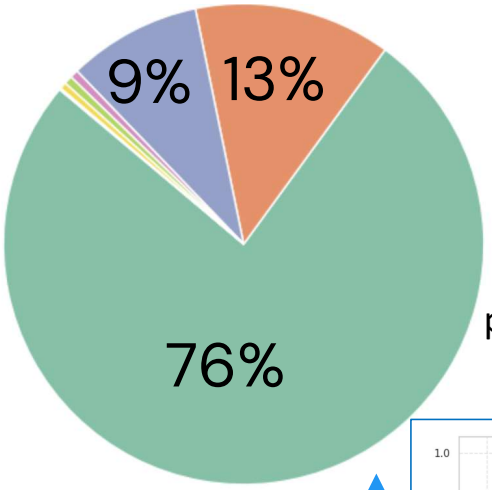
Cabin Count →

Parents/children and Sibling/Spouse

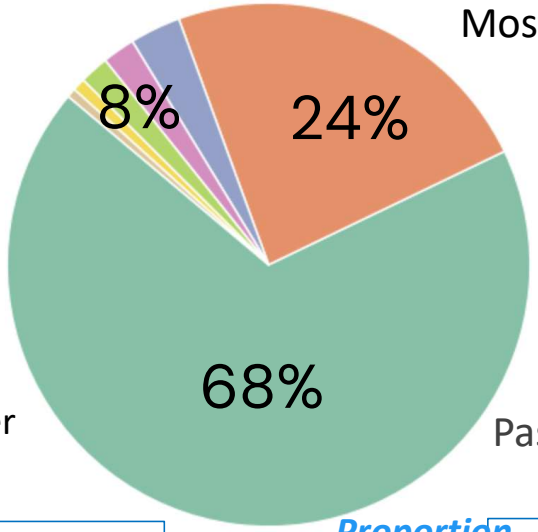
Analyze

Insights

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



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

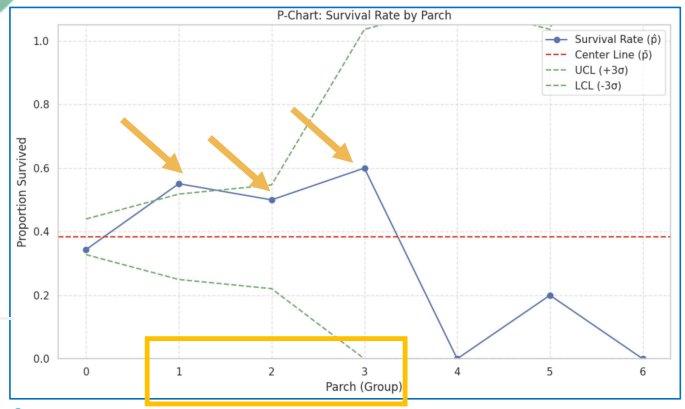


Key takeaways:
Most passengers traveled alone (solo), followed by those with 1-2 siblings/spouse (SibSp).

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

Proportion

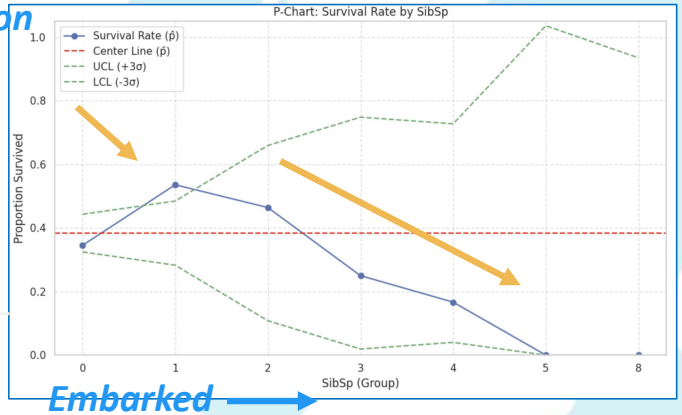
Cabin Group



Parch

Proportion

Embarked



Embarked

4

Exploratory Data Analysis

EDA – Insights and Feature Selection

Titanic Data

Analyze

Feature Engineering

Continuous

- Age
- Fare_log

Categorical

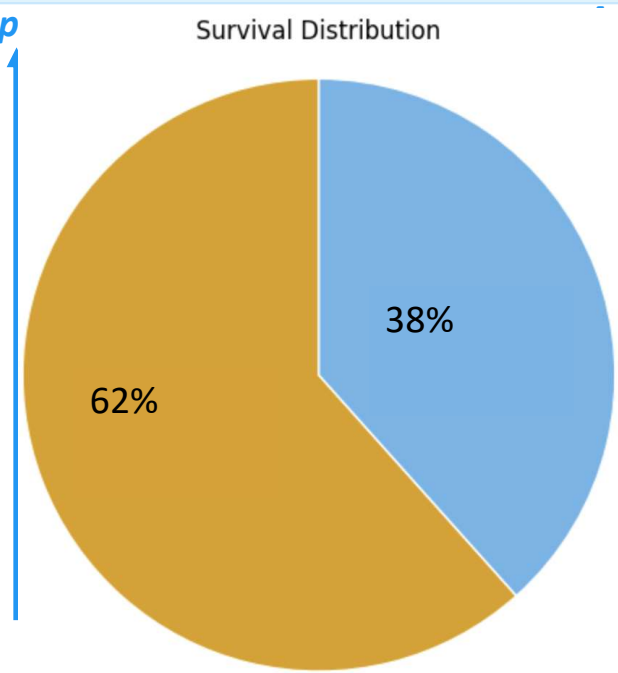
- Title
- Name
- Surname
- Sex
- PClass
- CabinGroup
- Embarked
- CabinNo
- Ticket

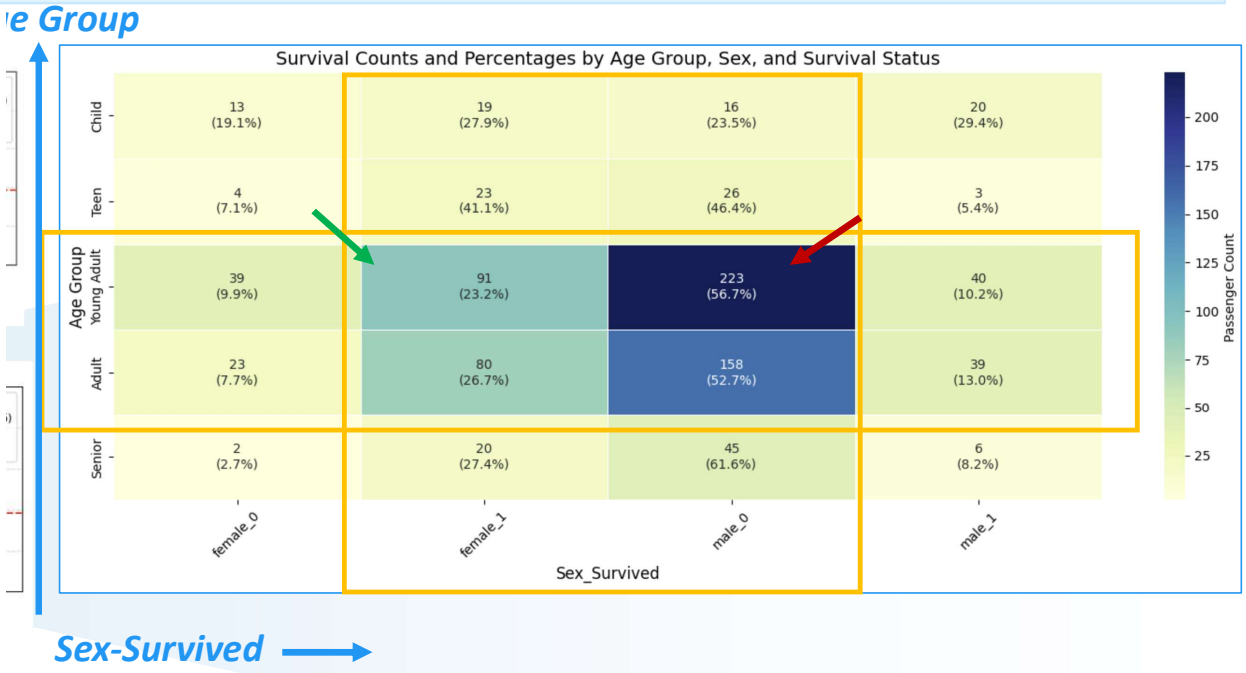
Count

- Parch
- SibSp
- CabinCnt

Next:
Passengers' segmentation (e.g. Female, 1st class, Adult, 94%)

Age Group, Sex and Survived



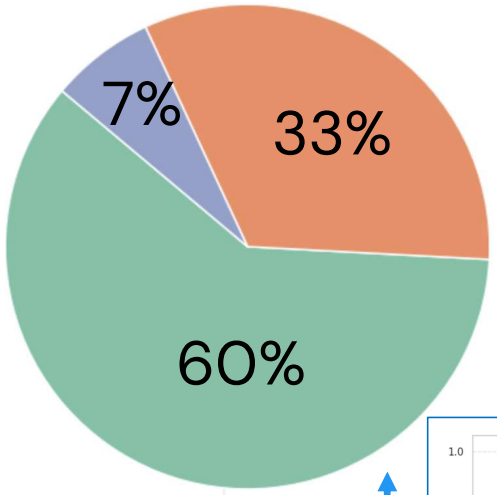


Age Group	Male SR	Female SR	Key Observations
Children (~70)	~30%	~30%	Highest male survival rate among all age groups
Teen (~55)	~5%	~40%	Teenage females had near-highest survival; males fared worse than children.
Young Adult (~390)	~10%	~25%	Males faced highest fatalities (>85%); female survival remained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

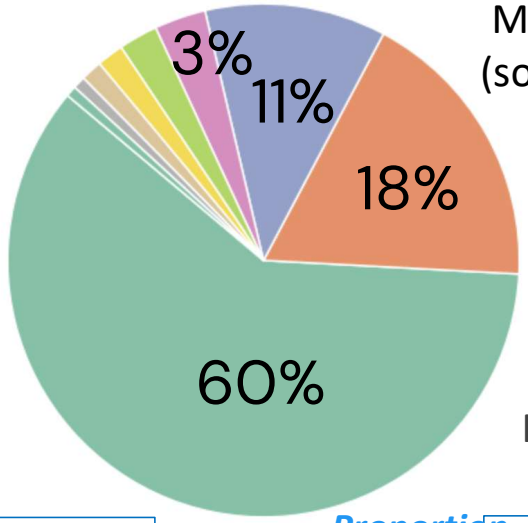
Family Group and Family Size

Analyze

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

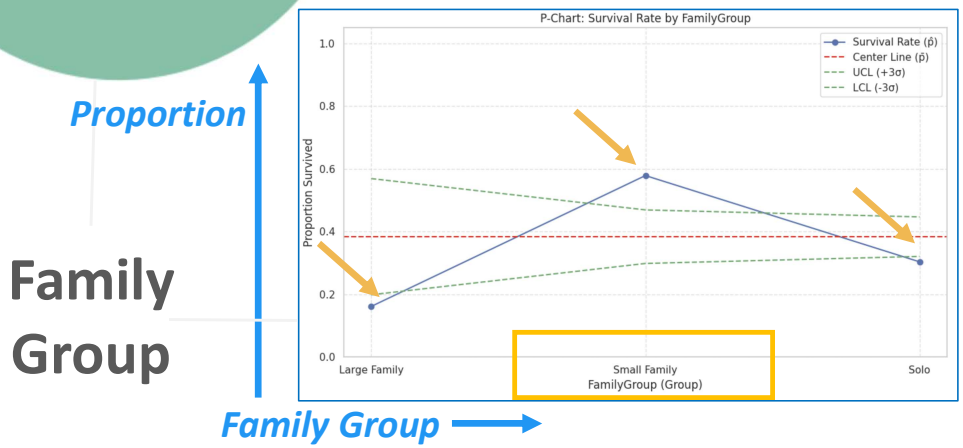


Insights

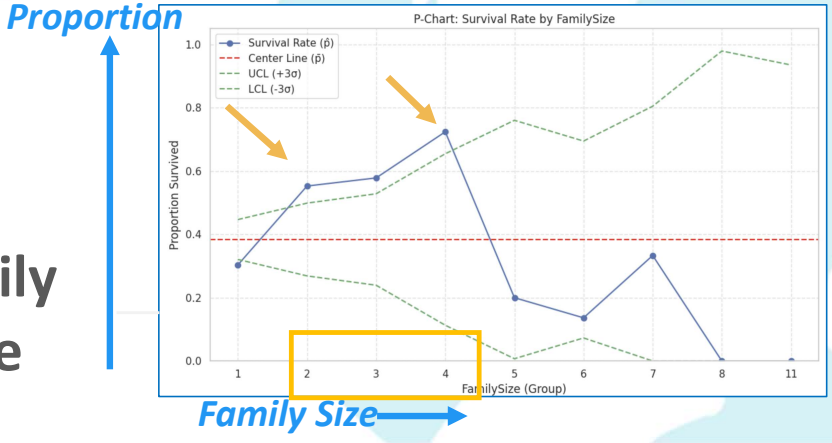


Key takeaways:
Most passengers with 1 family size (solo), followed by those with 2 to 4 sizes

Key takeaways:
Small family had higher survival rates



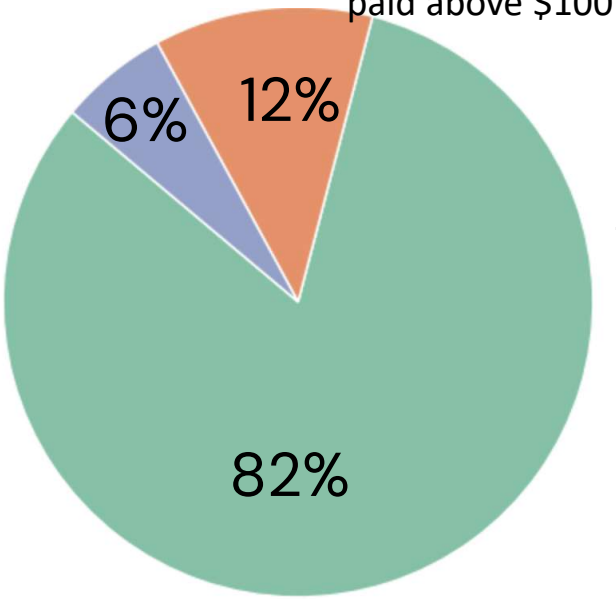
Key takeaways:
Family size from 2 to 4 had higher chance of survival



Fare Group

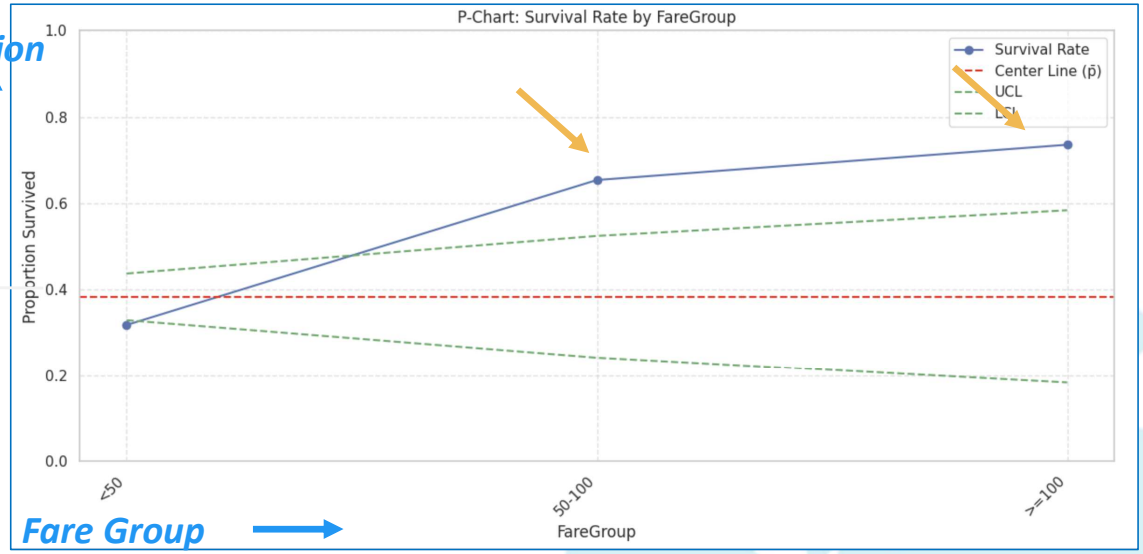
Analyze

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

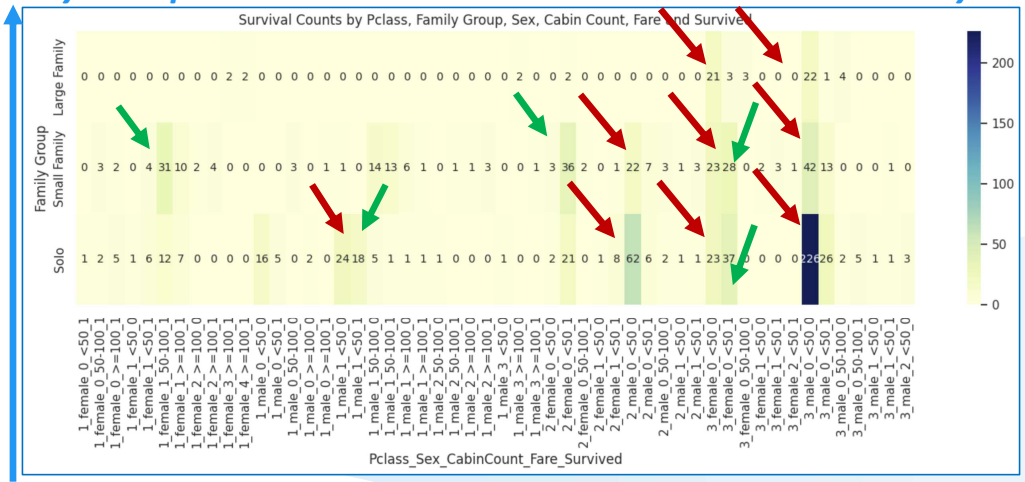
Insights



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

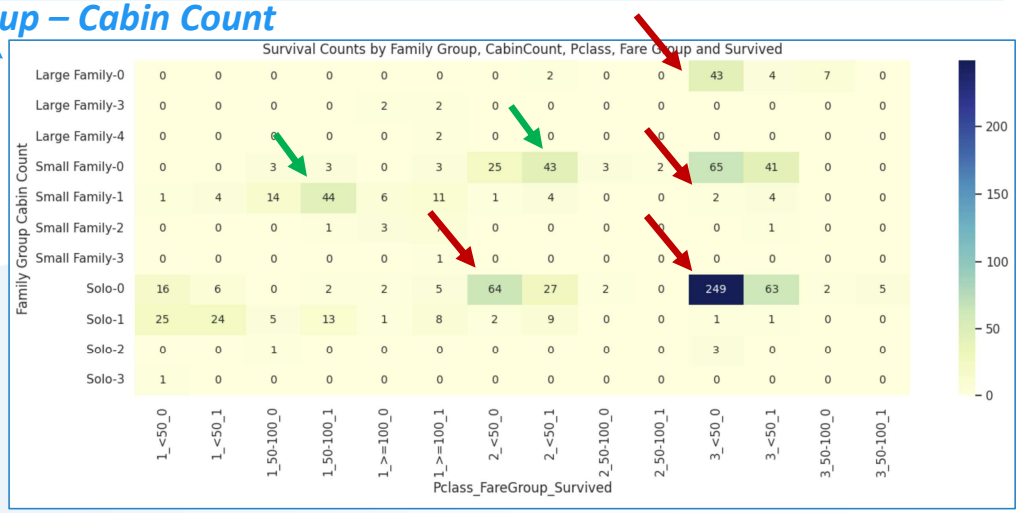
Family Group, Pclass, Cabin Count, Fare and Survived

Family Group



Pclass-Sex-Cabin Count-Fare- Survived →

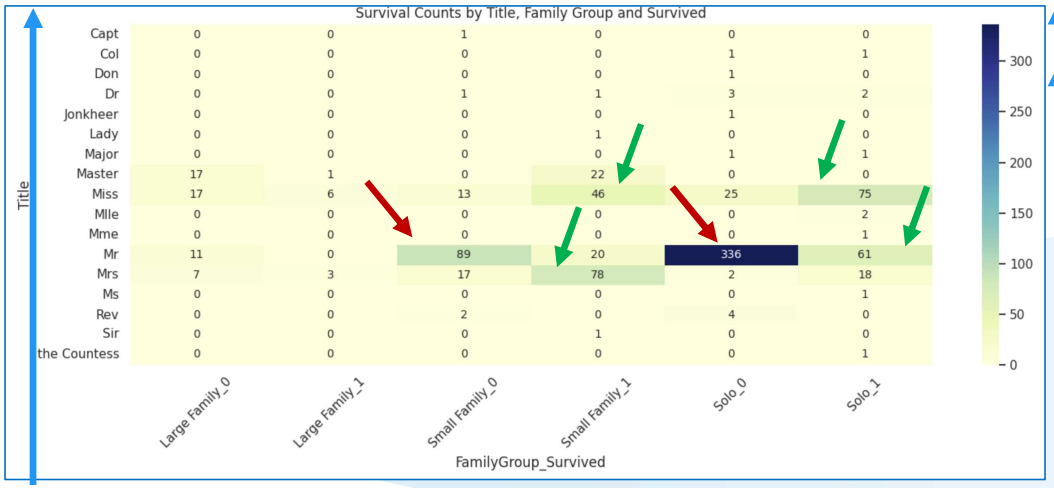
Family Group – Cabin Count



Pclass-Fare- Survived →

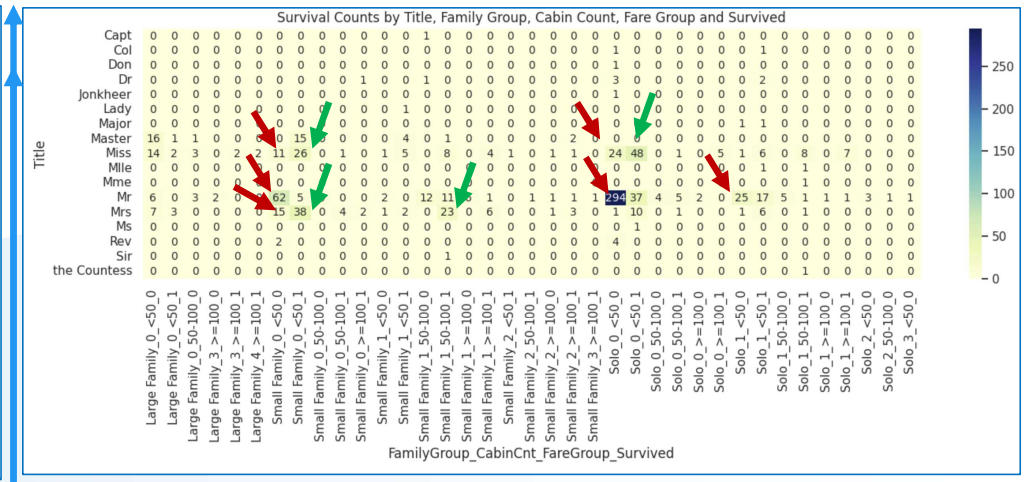
Family Group	Class	Cabin Count	Fare	Survival	Key Observations
Solo	3rd, 2nd	0	< $50	Mostly deceased	Highest fatality rate and second-highest male fatality
Small Family	3rd, 2nd	0	< $50	Mostly deceased	Severe male fatality group
Large Family	3rd	0	< $50	Mostly deceased	Severe male mortality
Solo & Small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

Title, Family Group, Cabin Count, Fare, Survived



Family Group-Fare-Surviced

Title



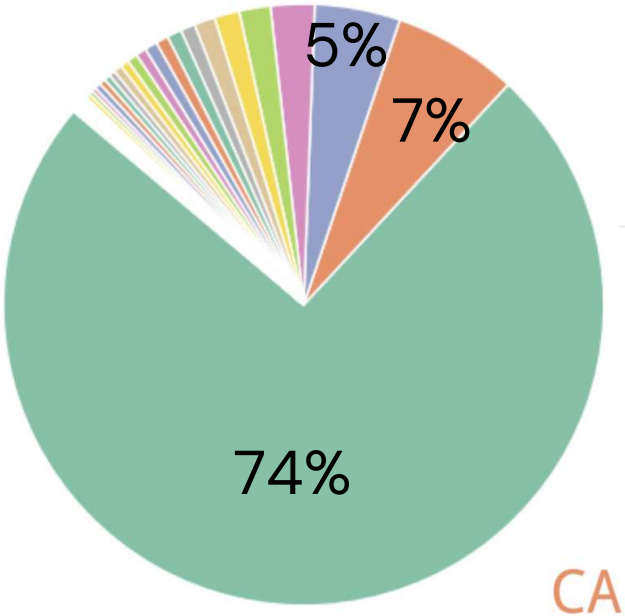
Family Group-Cabin Count-Surviced

Title	Family Group	Cabin Count	Fare	Survival	Key Observations
Mr (~520)	Solo	0	< $50	Mostly deceased	Highest fatality rate among all age groups
	Small Family	0	< $50	Mostly deceased	Second-highest male fatality group
	Solo	1	< $50	Mostly deceased	Severe male mortality
Miss (~180)/ Mrs(125)	Solo and Small Family	0	< $50	Mixed	High numbers in both of survivor/deceased categories
	Small Family	1	$50 - $100	Mostly survived	High female survivor rate (Mrs)

Ticket Group

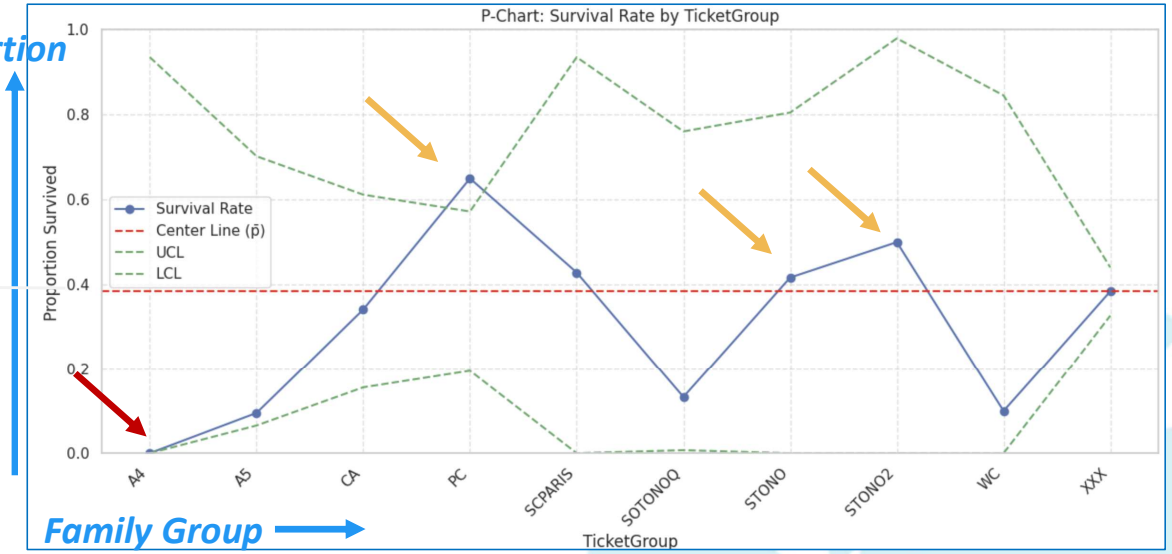
Analyze

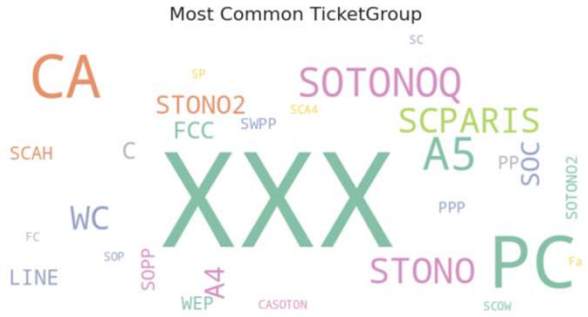
Key takeaways:
Most passenger's ticket only display numbers, followed by those with PC and CA



Ticket Group

Insights





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

Titanic Data

Analyze

Baseline – Features Selection

Continuous

- Age
- AgeGroup
- Fare_log
- FareGroup

Categorical

- Title
- PClass
- Surname
- CabinGroup
- Sex
- CabinNG
- Embarked
- TicketGroup

Count

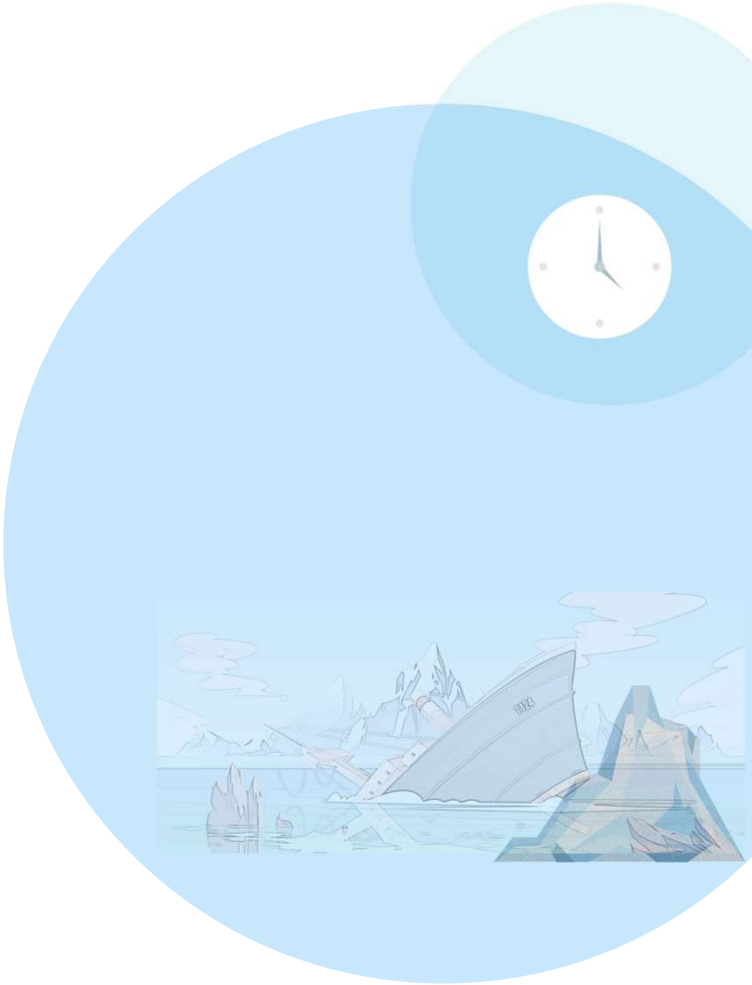
- Parch
- SibSp
- CabinCnt
- FamilySize

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

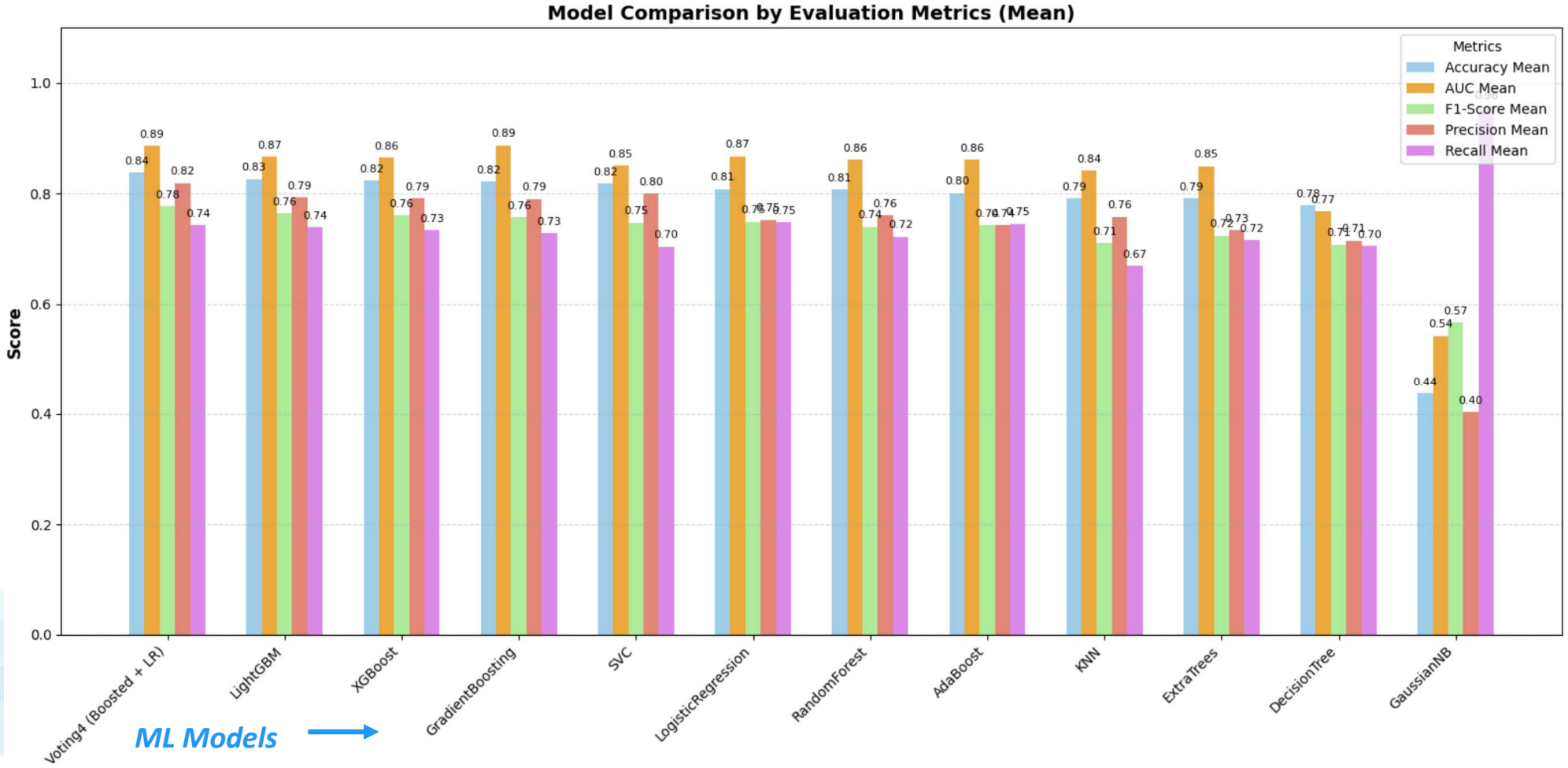
Y_train

Titanic Data

Construct

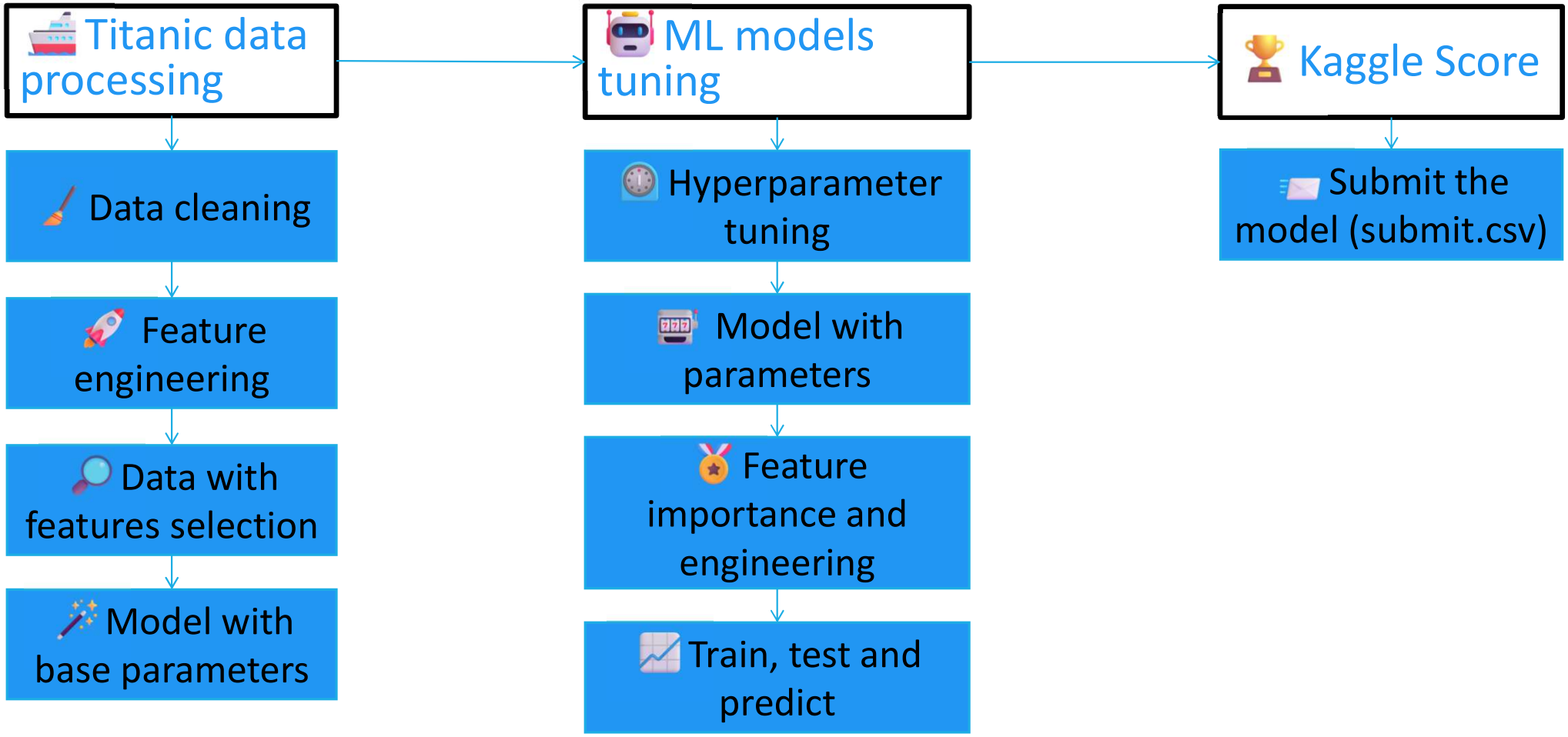
12 Models

Mean Score



ML Models →

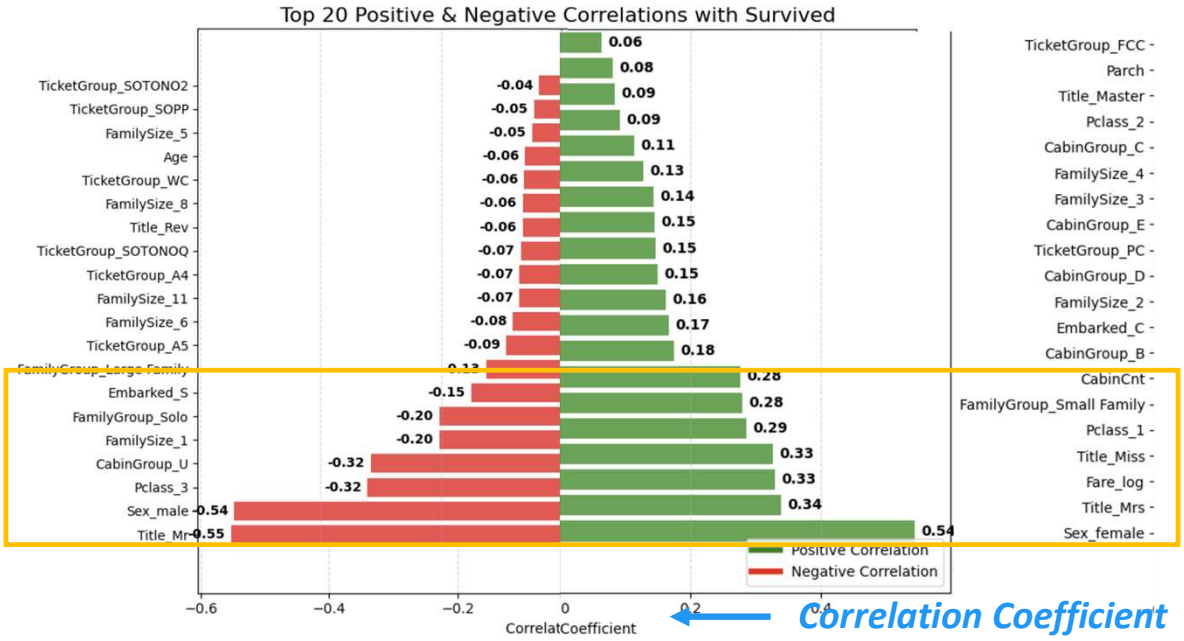
Flow Chart


Titanic data processing
Data cleaning
Feature engineering
Data with features selection
Model with base parameters
ML models tuning
Hyperparameter tuning
Model with parameters
Feature importance and engineering
Train, test and predict
Kaggle Score
Submit the model (submit.csv)

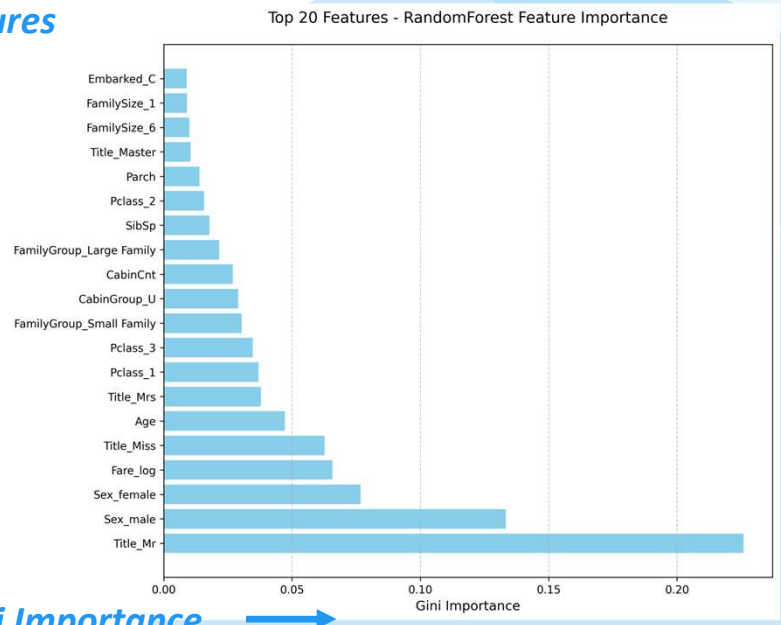
Titanic Data

Construct

Baseline – Features Selection



Features



Correlation Coefficient ***Gini Importance***

17 Features

- Title
- Fare_log
- CabinGroup
- Embarked
- Age
- Sex
- PClass
- CabinCnt
- FamilyGroup

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) — **Target**

83 Features (X) — **Y_train**

17 Features (X)

X_train_reduced | **X_test_reduced**

Titanic Data

Construct

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, treshold =0.03 ⬆️	83.2%	77.5%
Random Forest(10)	13 features (base), lightly imbalanced data, X_train_reduced_5, treshold =Pareto ⬆️	83.5%	77.3%
Random Forest(11)	21 features (base), lightly imbalanced data, X_train_reduced_6, treshold =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

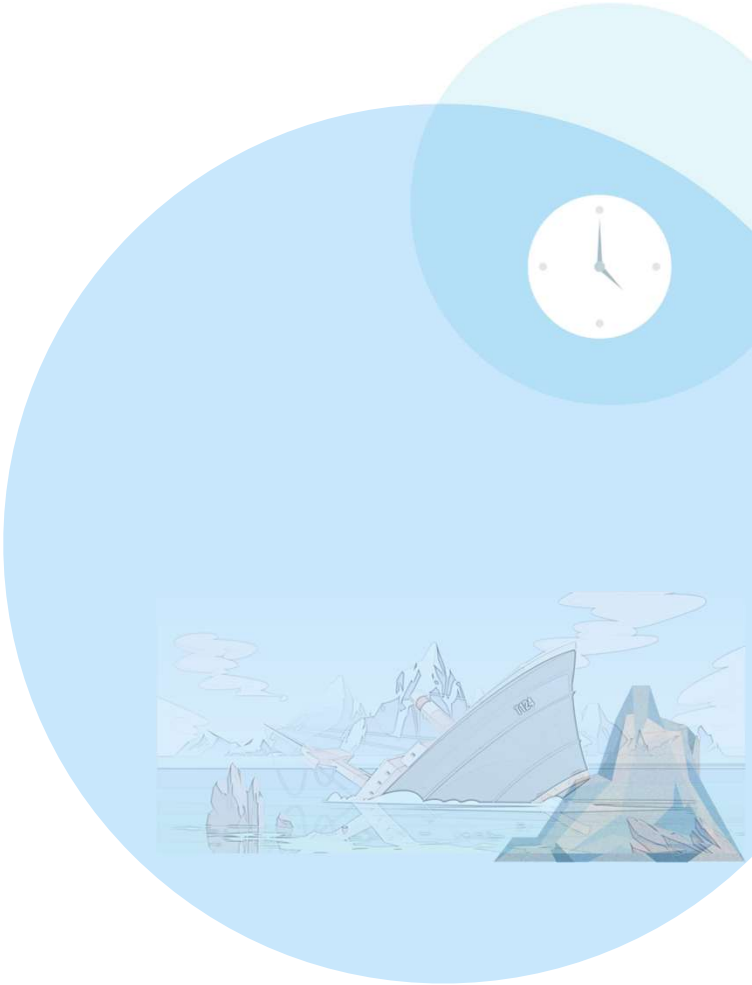
Construct

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

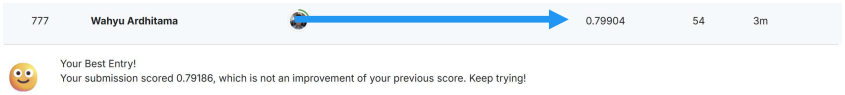
Conclusion

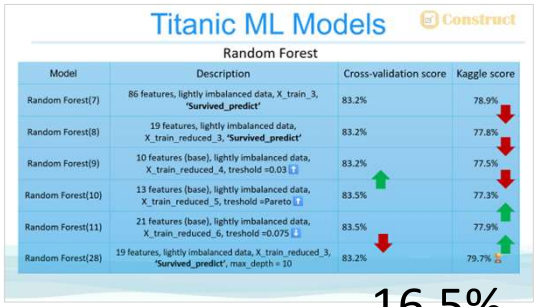


OKR Status

Execute

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 ❌

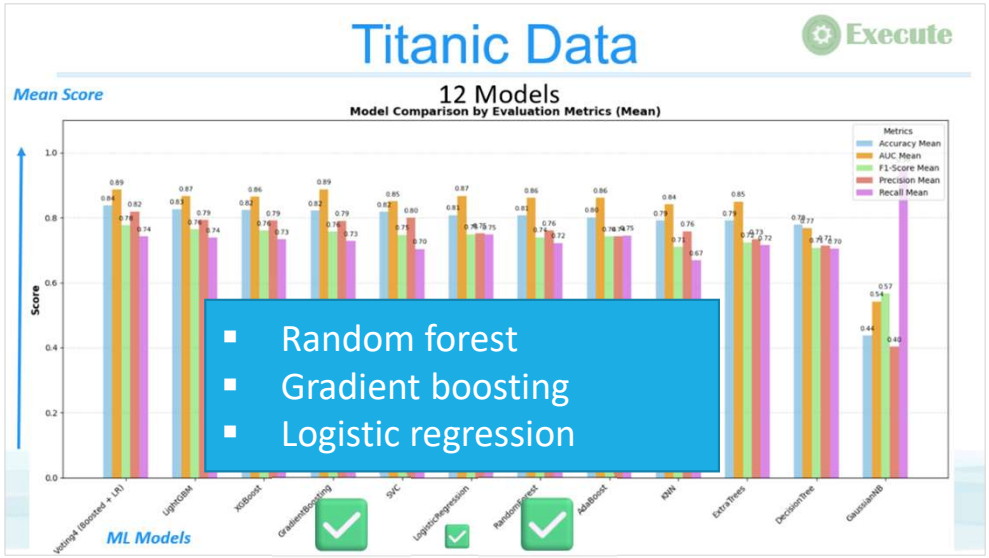




16.5%

OKR Status

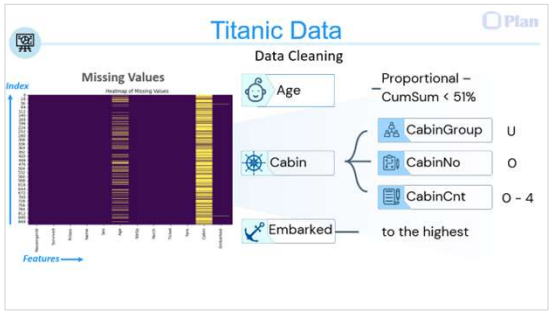
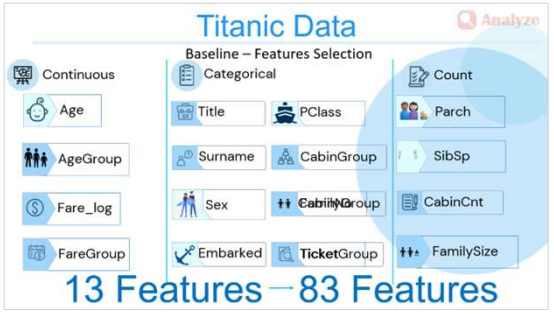
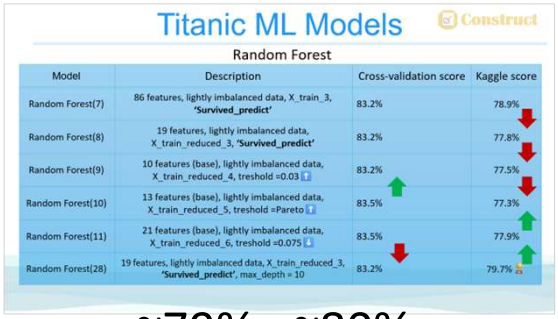
Execute



Benchmark and Evaluate Multiple ML Models	**Objective**
Next	**Status**
	Progress Score
May 5, 2025	**Deadline**
Train the promising machine learning models ✅	**Key Result 1**
Identify top 3 models based on accuracy and precision ✅	**Key Result 2**
Reduce runtime for model training by 20%	**Key Result 3**

OKR Status

Execute

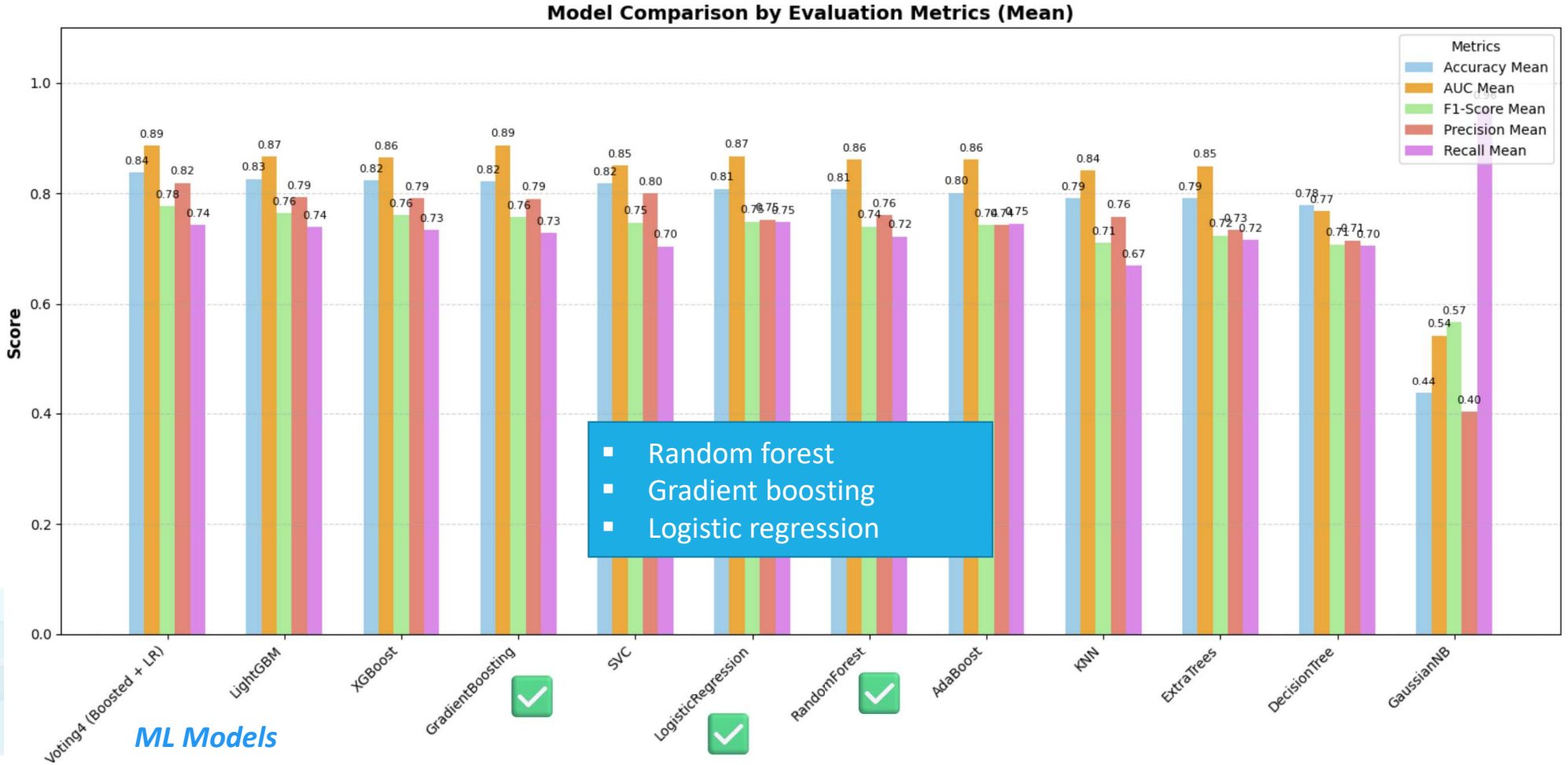
Objective		Enhance Feature Engineering and Data Quality
Status		**Next**
Progress Score		
Deadline		May 7, 2025
Key Result 1		Reduce missing data to <1% across all relevant features
Key Result 2		Create at least 5 new meaningful engineered features
Key Result 3	~79% - ~80%	Improve model performance by at least 5% after feature engineering

Titanic Data

Execute

12 Models

Mean Score



- Random forest
- Gradient boosting
- Logistic regression

ML Models

Data Project Architecture

Execute

Project Management + Business Management

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

Development

Plan

1st step
Import the relevant library and data

2nd step
Data cleaning

Analyze

1st step
SPC

2nd step
Aggregation and visualization

3rd step
Feature engineering

Construct

1st step
Construct the model

2nd step
Feature importance

3rd step
Save models

Deployment

Power BI

1st step
Developing

2nd step
Publishing

Streamlit

1st step
Developing

2nd step
Running

Key takeaways:
Multiple roles vs multi tasking

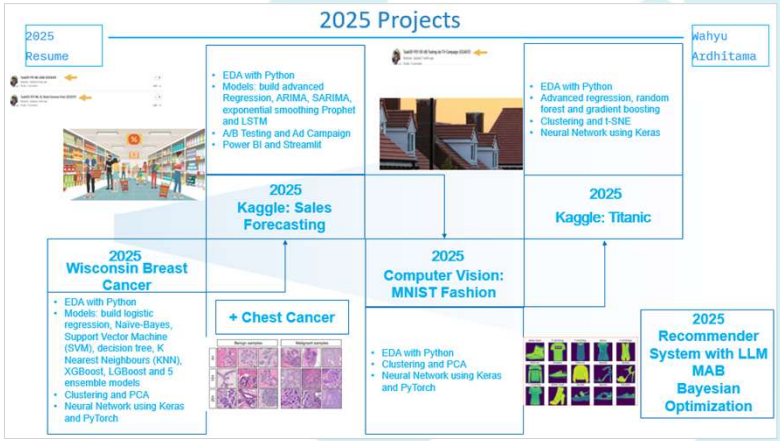
Next Steps

Improvements:

- Continue refining the Random Forest model to achieve ≥ 80% accuracy
- Expand model testing beyond random forest to include:
  - Gradient boosting: capture complex non-linear relationships
  - Logistic regression: offers interpretable results and serves as a baseline model
- Enhanced feature engineering, e.g., using Fare_group or incorporating feature importance from other models (e.g., SHAP values).
- Implement a Voting Classifier that combines Random Forest, Gradient Boosting, and Logistic Regression for improved overall performance

Next projects:

- Wisconsin breast cancer + chest cancer (image data)
- Kaggle: sales forecasting
- Computer vision: MNIST – fashion
- Kaggle: house price



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