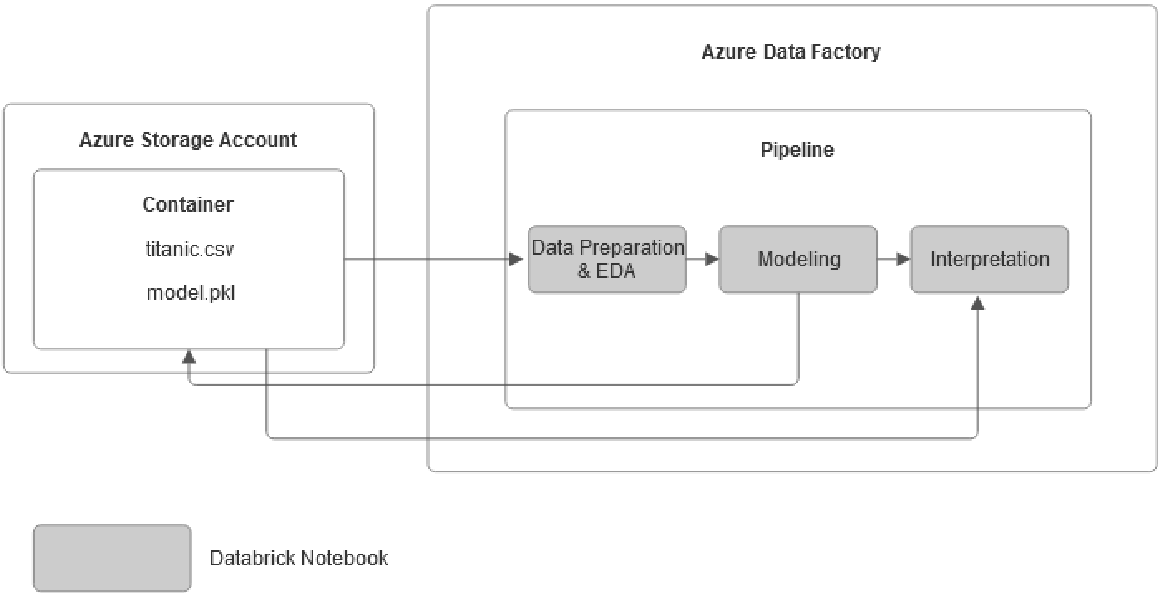
Project Topic: Titanic Survival Prediction from Demographic Factors

Project Overview:

Titanic, an ocean liner that sank on 15 April 1912 as a result of striking an iceberg, was one of the most famous tragedies in modern history. Predicting the passengers based on demographic information can significantly provide insights into how different ages, genders, social statuses are treated differently in the face of disaster and shed light on to the prejudice and social perception of different people during the time of the incident. This project involves building different classification models to output the most likely outcome for a particular passenger given their demographic information and deploy real-time prediction based on user input.

Implementation Steps:

1. **Data Collection:**
    - Connect to the data sources through Azure storage account container, in the format of csv
2. **Data Cleaning and Preprocessing:**
    - Use libraries like Pandas and NumPy for data manipulation.
    - Handle missing values using imputation techniques
3. **Exploratory Data Analysis:**
    - Utilize visualization tools like Matplotlib and Seaborn to visually present the hidden relationships within data. Specifically, use comparative histograms and bar charts to compare the survival outcome within different demographic strata.
    - Identify key factors in correlation with survival through hypothesis testing. Specifically, implement t-test for numerical factors and chi-square test of independence for categorical factors.
    - Analyze correlations between factors using Pearson correlation coefficients, scatterplots, bar charts, VIF analysis to uncover potential multi-collinearity.
4. **Feature Engineering:**
    - Create new features using domain knowledge and interaction terms.
5. **Model Building:**
    - Implement models using Scikit-Learn.
    - Experiment with different models including logistic regression, polynomial logistic regression, ridge logistic regression, decision tree, random forest, gradient boosting classifier.
    - Use KFold and GridSearchCV for hyperparameter tuning.
    - Use train-test split for model selection.
6. **Deployment:**
    - Develop notebooks in Databricks
    - Develop an Azure Data Factory pipeline to automate notebook runs
    - Save best model as model artifact in the form of pkl
    - Deploy Flask for user input and real-time predictions
7. **Monitoring and Maintenance:**
    - Regularly update the model with new data and monitor performance.



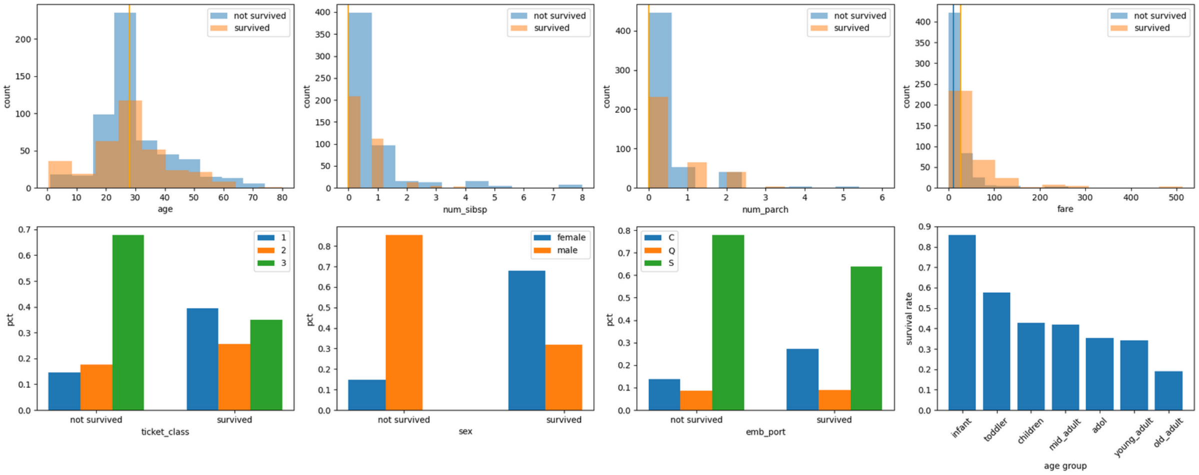
Expected Outcomes:

- Potential demographic factors contributing to titanic survival
- Model artifact of the model that classifies titanic survival with highest accuracy
- Webpage that imports model artifact and makes real-time prediction based on inputs from user prompts

Tools and Technologies:

- **Programming Languages:** Python, SQL
- **Libraries:** Pandas, NumPy, Scipy, Sciksit-Learn, Matplotlib, Seaborn
- **Deployment:** Flask
- **Cloud Platforms:** Azure

Analysis Details:



Factor	Distribution	t-stat	P-value
age	mean age for survived: 28.16 mean age for not survived 30.0	-2.08	0.037
num_sibsp	mean num_sibsp for survived: 0.48 mean num_sibsp for not survived 1.0	-1.01	0.311
num_parch	mean num_parch for survived: 0.47 mean num_parch for not survived 0.0	2.485	0.013
fare	mean fare for survived: 48.21 mean fare for not survived 22.0	7.863	0.00

Factor	Distribution (pct of not survived vs pct of survived)	Chi-square Stat	P-value
ticket_class	First class: 1: 0.15 vs. 0.39 Second class: 0.18 vs. 0.26 Third class: 0.68 vs. 0.35	100.980	0.00
sex	Female: 0.15 vs. 0.68 Male: 0.85 vs. 0.32	258.427	0.00
emb_port	Cherbourg: 0.14 vs. 0.27 Queenstown: 0.09 vs. 0.09	26.489	0.00

	Southampton: 0.78 vs. 0.64		

Key Insights:

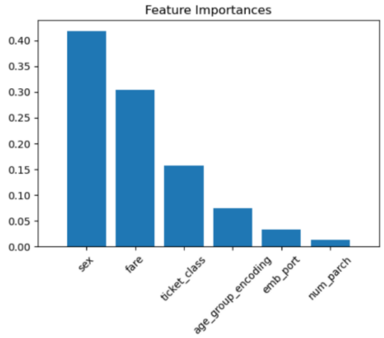
- Demographic factors play a role in surviving the titanic shipwreck.
- Large disparity of survival is observed among different fares, ticket class and sex
- Disparity of survival is observed among different embarkment ports, age, and number of parents/children.
- Passengers that are female/obtaining first class tickets/young/having parents or children showcase higher chance of survival as opposed to other groups.
- The disparity of survival observed among embarkment ports is partially contributing to their associated fare and ticket class with rank Cherbourg > Southampton > Queenstown.

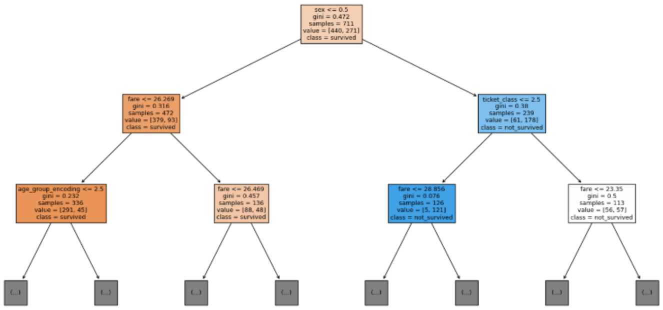
Model Performance:

Model	Train Accuracy	Test Accuracy
Logistic Regression	0.806	0.781
Ridge Logistic Regression	0.788	0.792
Polynomial Logistic Regression	0.947	0.781
Random Forest	0.947	0.792
Gradient Boosting Classifier	0.924	0.831

Model Insights

Based on the feature importances of the best model – gradient boosting classifier - sex, ticket fare, and ticket class are the most important features in predicting the survival, followed by age group. Similarly, top features of the decision tree are dominated by sex, fare, and ticket class. These observations align with the principle of "women and children first" and "first class first."


Feature Importances
sex
fare
ticket_class
age_group_encoding
emb_port
num_parch


sex <= 0.5
gini = 0.472
samples = 711
value = [440, 271]
class = survived
fare <= 26.269
gini = 0.316
samples = 472
value = [379, 93]
class = survived
ticket_class <= 2.5
gini = 0.38
samples = 239
value = [61, 178]
class = not_survived
age_group_encoding <= 2.5
gini = 0.232
samples = 336
value = [291, 45]
class = survived
fare <= 26.469
gini = 0.457
samples = 136
value = [88, 48]
class = survived
fare <= 28.856
gini = 0.076
samples = 126
value = [5, 121]
class = not_survived
fare <= 23.35
gini = 0.5
samples = 113
value = [56, 57]
class = not_survived
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)