PROJECT 1 : Titanic - Predicting Survival and Telling a Data Story

Deliverables
Instructions
Recommendations
Timeline
PROJECT : Titanic
Step 1: Data Cleaning and Preparation
Step 2: Feature Transformation and Engineering
Step 3: Visualization and Exploration (with `ggplot2`)
Step 4: Modeling and Prediction (Data Science)

▼ Deliverables

- An `.Rmd` file.
- The corresponding knitted HTML file.
- The presentation slides (submitted as a PDF).

▼ Instructions

For the Rmd/HTML report, explanations must be provided for each step. The layout must be professional and "business-ready," paying close attention to Markdown formatting, the use of a CSS file for styling, and the inclusion of a Table of Contents (TOC).

Regarding the presentation, the total speaking time is 10 minutes (to be shared), followed by a 5-minute Q&A session. The presentation support must be slides in PPTX format; while there is no maximum slide count, there should be at least 6 slides at the minimum. The slides must be submitted in PDF format.

It would be appreciated if you use a Notion/Git or any project tool ; that will give you bonus points.

▼ Recommendations

Pay special attention to the polish and formatting.. The presentation should concentrate on a quick overview of the case study, its analyses (focusing on the **results**, not the technical operations), its visualizations, and its conclusions, including different avenues for planned or future improvements.

▼ Timeline

- Finalisation of the groups : Saturday 18/10 23H59
- Start of the project : Sunday 19/10 00H00
- Next session when you can ask any questions : Wednesday 22/10
- Time limit of the submission of project files (Rmd/HTML) : Monday 10/11 at 23h00
- Time limit of the submission of presentation file (pdf) : Wednesday 12/11 at 7h00
- Presentation of the project : Wednesday 12/11

For each student , you'll have to submit the files : just submit one version of it on Junia Learning ; then for each group, you'll tell me which member has the latest version of the project.

PROJECT : Titanic

Project Goal : The sinking of the Titanic is a well-known historical tragedy. Your mission is to use passenger data to build a predictive model of survival. Beyond just prediction, you must conduct an in-depth analysis to understand which factors, from the obvious to the subtle, truly influenced a passenger's chances of survival. Your final analysis should tell a compelling story, supported by visualizations and your model's findings.

Source : https://www.kaggle.com/c/titanic/data

Step 1: Data Cleaning and Preparation

- **Handling Missing Ages :** Impute these values using the median age, but do so group-wise by passenger class and sex for a more accurate estimate.

- `Cabin` **Column :** Transform it into a new binary variable called `HasCabin` (see what is the meaning behind the NA values).
- `Embarked` column
- **Verifying Data Types :** Ensure that categorical variables are correctly encoded as "factors" in R.
- Check the other columns

Step 2: Feature Transformation and Engineering

- **Creating Family Size:** Combine the `SibSp` (siblings/spouses) and `Parch` (parents/children) columns to create a new numerical feature called `FamilySize.`
- **Extracting Titles :** Create a new feature `Title`. Then, group the rarest titles into an "Other" category.
- **Creating Age Groups :** Transform the numerical `Age` variable into an ordered categorical variable (`AgeGroup`) by creating relevant bins (e.g., 'Child', 'Adolescent', 'Adult', 'Senior').
- **Analyzing Fare Per Person:** Some tickets (`Ticket`) were purchased for multiple people. Create a new feature `FarePerPerson` by dividing the `Fare` by `FamilySize` for passengers traveling in groups.
- Create any additional columns that you feel is missing.

Step 3: Visualization and Exploration (with `ggplot2`)

- **Survival Rate by Sex and Class:** Create a bar chart showing the count of survivors and non-survivors, faceted by `Sex` and `Pclass.`
- **Age Distribution and Survival:** Compare the age distributions between survivors and non-survivors.
- **Relationship between Fare and Survival: V**isualize the distribution of `Fare` based on survival status, separated by `Pclass`.

- **Impact of Family Size:** Create a visualisation showing the survival rate for each `FamilySize` to determine if people traveling alone or in large families had different outcomes.
- Add 2 pertinent visualisations of your liking.

Step 4: Modeling and Prediction (Data Science)

- **Baseline Model:** Train a first logistic regression model to predict the `Survived` variable using only the basic features (`Pclass`, `Sex`, `Age`).
- **Improved Model:** Train a second logistic regression model that includes the new features you engineered.
- **Decision Tree Model:** Train a decision tree model (using the `rpart` package) to get a visual representation of the decision rules that lead to a survival prediction.
- **Model Evaluation:** Compare the performance of your models using a confusion matrix to calculate accuracy and analyze the feature importance for each model.