

Standard Data Science Template

Project Template

Analytics Team

2024-01-01

Table of contents

1	Introduction	2
2	Problem Statement	2
3	Exploratory Data Analysis	2
4	Feature Engineering	5
5	Model Selection	6
6	Analysis of Feature Relationships	6
7	Model Tuning	6
8	Model Validation and Testing	6
9	Results	6
10	Conclusion	6
11	Appendix	6
12	Example PDF Usage & Formatting	6

1 Introduction

Describe the dataset.

2 Problem Statement

Describe the problem. What are we trying to predict? Is there a baseline to measure against? Does prediction bring value? “So what?”

3 Exploratory Data Analysis

1. Profile the dataset.

Check for correct data types, nulls, uniqueness, granularity and top value counts.

Table 1: Quality Check of All Fields

	Data Type	Unique Count	Percent Null	Mode	Mode % of total
Survived	Integer	2	0%	0	62%
Sex	String	2	0%	male	65%
Pclass	Integer	3	0%	3	55%
Embarked	String	3	0%	S	72%
SibSp	Integer	7	0%	0	68%
Parch	Integer	7	0%	0	76%
Age	Float	88	20%	24.0	4%
Cabin	String	147	77%	B96 B98	2%
Fare	Float	248	0%	8.05	5%
Ticket	String	681	0%	347082	1%
Name	String	891	0%	Braund, Mr. Owen Harris	0%

- 2) Compute descriptive statistics of numeric fields.

Table 2: Descriptive Statistics of Numeric Fields

	Min	Mean	Median	Max	Standard Dev	Kurtosis
Survived	0	0.4	0.0	1	0.5	-1.8
Pclass	1	2.3	3.0	3	0.8	-1.3
Age	0	29.7	28.0	80	14.5	0.2
SibSp	0	0.5	0.0	8	1.1	17.9
Parch	0	0.4	0.0	6	0.8	9.8
Fare	0	32.2	14.5	512	49.7	33.4

3) Explore dependent variable

Not necessary, as it is binary. Accomplished above.

4) Visualize independent variables.

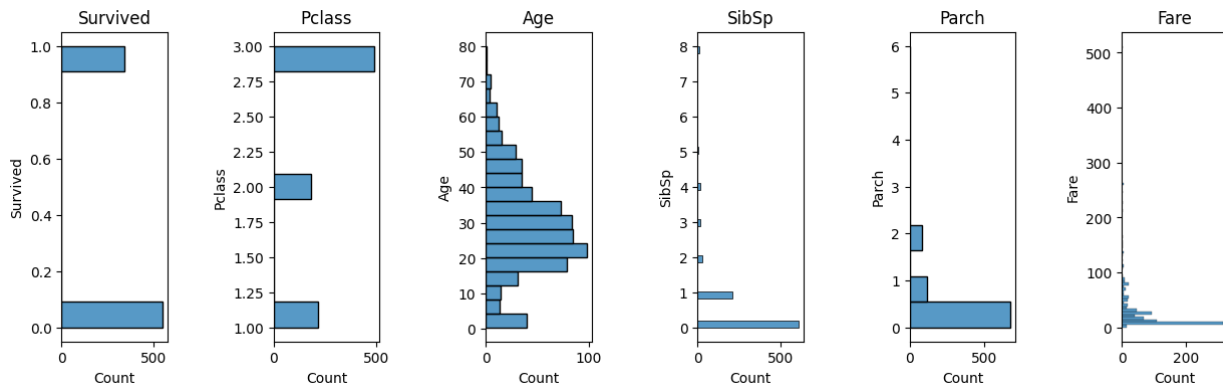


Figure 1: Distribution of Numeric Fields

5) Explore relationship independent variables have on dependent variable.

- Correlations
- Predictive Power Scores

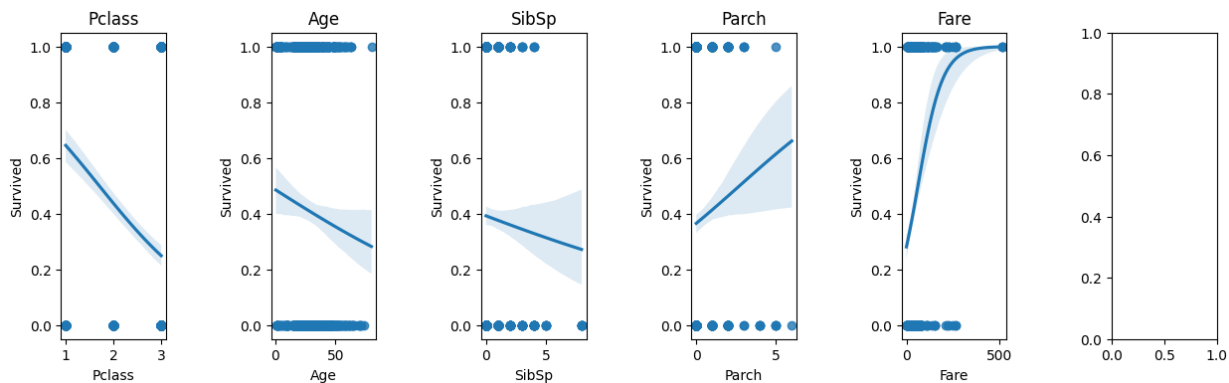


Figure 2: Relationship Between Numeric Fields and Target

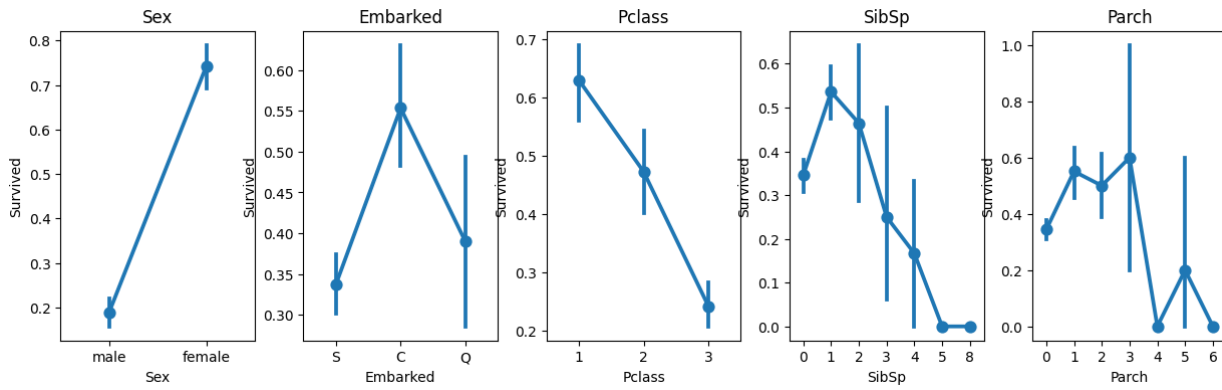


Figure 3: Relationship between Categorical Fields and Target

Unable to display output for mime type(s): application/vnd.plotly.v1+json

Parallel categories plot with respect to target

Unable to display output for mime type(s): application/vnd.plotly.v1+json

Parallel coordinates plot with respect to target

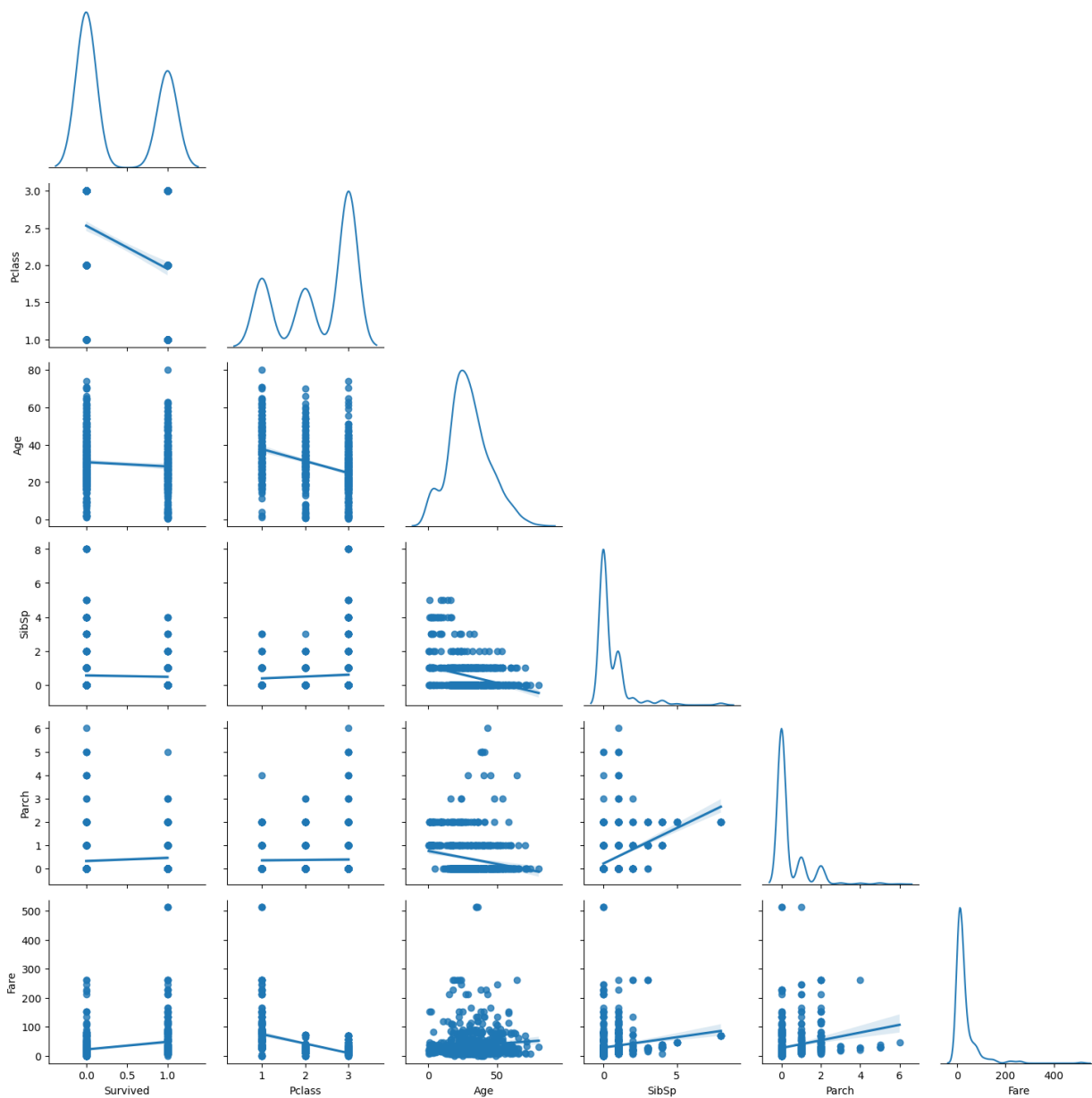


Figure 4: Relationships Between Numeric Fields

4 Feature Engineering

Should we remove outliers? manually impute nulls? handle high-value-count categories? handle date or time columns? convert data types?

Select a specific set of features and optionally re-name them.

5 Model Selection

Run autoML to train the model. This can be for either regression or classification, but the focus here will be for regression and clustering

6 Analysis of Feature Relationships

Calculate shap values for the model and visualize them with respect to each feature

Look at pair plot and parallel coordinates (plotly or hiplot)

7 Model Tuning

8 Model Validation and Testing

9 Results

10 Conclusion

11 Appendix

12 Example PDF Usage & Formatting

- Jupyter cell behaviour:
 - the following code can be added to the top of a cell to change how it renders in the PDF
 - toggle output
 - * `#| output: false`
 - * `#| output: true`
 - Toggle code in output
 - * `#| echo: false (default)`
 - * `#| echo: true`
- add captions to an output
- `#| fig-cap:` caption for plot
- `#| tbl-cap:` caption for table

- [more documentation here](#)
- Plotly visualization shortcuts code shortcuts:
 - px-fig - Create over 30 types of statistical and scientific graphics figures.
 - px-update - Update layout and data trace styling of existing figure.
 - px-args - Select arguments from lists of options to modify figure styling.