**Лабораторная работа №1.2. Построение и отбор признаков.**

**Введение**

Машинное обучение – это не просто подбор правильных параметров для модели. Рабочие процессы ML зависят от множества других аспектов, включая построение и отбор признаков. Разберемся, для чего нам нужны признаки, а также изучим особенности реализации техники feature engineering.

Признак “Cabin” сложен и требует дополнительного изучения. Большое количество значений “Cabin” отсутствует, но сам признак нельзя полностью игнорировать, потому что некоторые значения могут оказывать большое влияние на показатель выживаемости. Оказывается, первая буква значения Cabin — это палубы, в которых расположены каюты. Эти палубы в основном были разделены для одного пассажирского класса, но некоторые из них использовались несколькими пассажирскими классами одновременно.

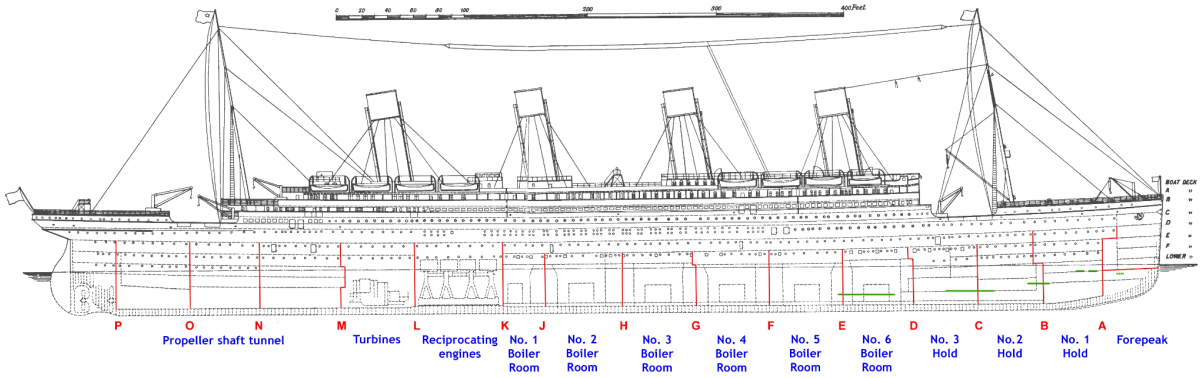


Рисунок 1. Схема палуб корабля Титаник

- На лодочной палубе было 6 типов палуб, помеченных как T, U, W, X, Y, Z, но в наборе данных присутствует только каюта T.
- Палубы A, B и C предназначались только для пассажиров 1-го класса.
- Палубы D и E были для всех классов
- Палубы F и G предназначались для пассажиров 2-го и 3-го классов.
- От A до G расстояние до лестницы увеличивается, что может быть фактором выживания.

Создадим столбец палуба Deck, содержащий первую букву признака “Cabin” (Значение M для отсутствующих значений кают). Выполнить это можно либо с помощью метода .apply() и lambda функции, либо же используя .loc[].

```
df_all['Deck'] = #Добавьте свой код сюда
```

Итоговое распределение пассажиров по палубам должно выглядеть следующим образом:

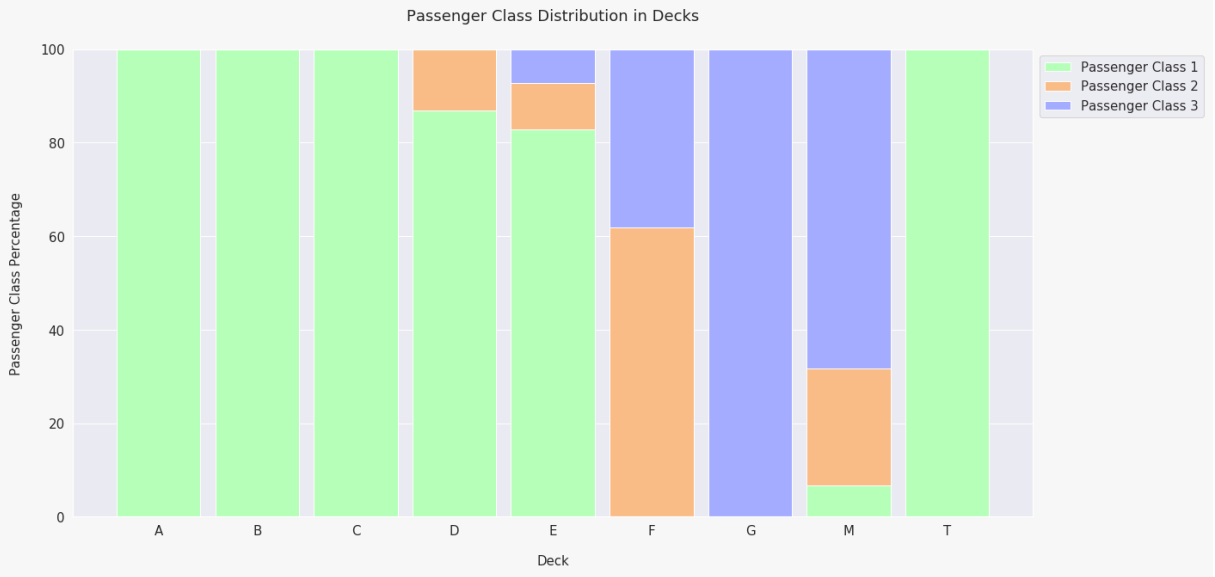


Рисунок 2. Распределение пассажиров по палубам

- 100% палуб A, B и C занимают пассажиры 1-го класса.
- На палубе D 87% пассажиров первого класса и 13% пассажиров второго класса.
- На палубе E 83% пассажиров 1-го класса, 10% 2-го класса и 7% пассажиров 3-го класса.
- На палубе F 62% пассажиров 2-го класса и 38% пассажиров 3-го класса.
- 100% палубы G — пассажиры 3-го класса.
- На шлюпочной палубе в каюте T находится один человек, он является пассажиром 1-го класса. Пассажир с этой палубы больше всего по своим признакам похож на пассажиров с палубы A, поэтому его можно отнести к пассажиром с палубы A.
- Значение M можно принять, как отдельную палубу, т.к. истинное значение палубы восстановить не получится.

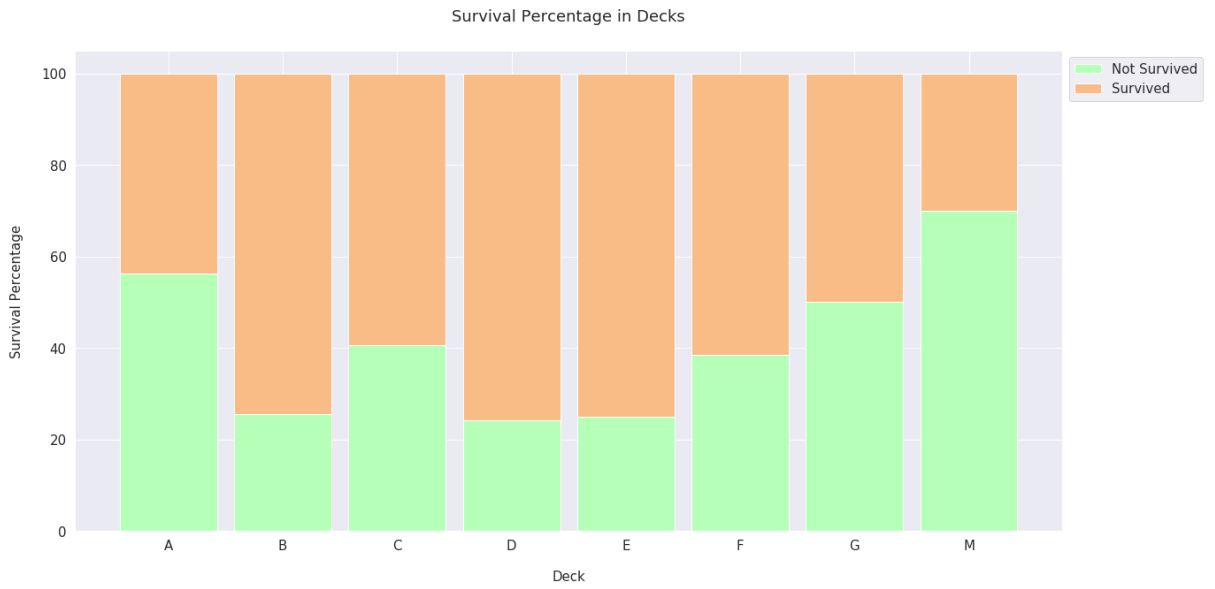

Рисунок 3. Выживаемость пассажиров в зависимости от занимаемой палубы.

Каждая палуба имеет разные показатели выживаемости, и эту информацию нельзя упускать из виду. Палубы B, C, D и E имеют самые высокие показатели выживаемости. Эти палубы в основном заняты пассажирами 1-го класса. M имеет самый низкий процент выживаемости, который в основном занят пассажирами 2-го и 3-го классов. Таким образом, каюты, используемые пассажирами 1-го класса, имеют более высокие показатели выживаемости, чем каюты, используемые пассажирами 2-го и 3-го классов. У пассажиров на палубе M (отсутствующие значения каюты) самая низкая выживаемость, это уникальная группа с общими характеристиками. На основе сходства некоторых признаков можно произвести их группировку:

- Палубы A, B и C объединить в палубу ABC, потому что на всех них находятся только пассажиры 1-го класса.
- Палубы D и E помечены как DE, потому что обе они имеют одинаковое распределение пассажирских классов и одинаковую выживаемость.
- Палубы F и G помечены как FG по той же причине, что и выше.
- Палубу M не нужно группировать с другими палубами, потому что она сильно отличается от других и имеет самую низкую выживаемость.

**Binning (Квантование признаков)**

Проблема работы с необработанными, непрерывными числовыми признаками заключается в том, что часто распределение значений в этих признаках будет искажено. Это означает, что некоторые значения будут встречаться довольно часто, а некоторые будут довольно редко. Помимо этого, существует также другая проблема изменяющегося диапазона значений для этих признаков. Например, количество просмотров определенных музыкальных видео может быть чрезмерно большим, а некоторые могут быть очень маленькими. Непосредственное использование этих признаков может вызвать множество проблем и негативно повлиять на работу модели. Следовательно, есть стратегии для борьбы с этим, которые включают в себя биннинг и преобразования.

Биннинг, также известный как квантование, используется для преобразования непрерывных числовых признаков в дискретные (категории).

Воспользоваться квантованием можно, например для признаков “Fare” и “Age”, для этого можно использовать функцию `pd.qcut`. Примеры преобразованных признаков приведены на рис. 4-5.

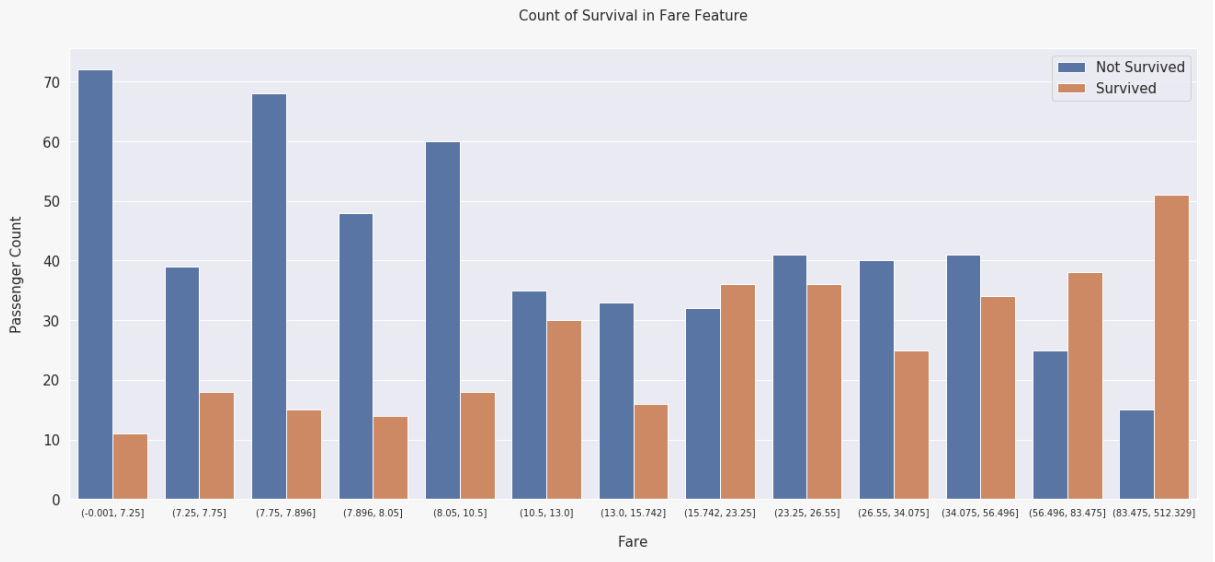


Рисунок 4. Количество выживших, приходящихся на бины Fare

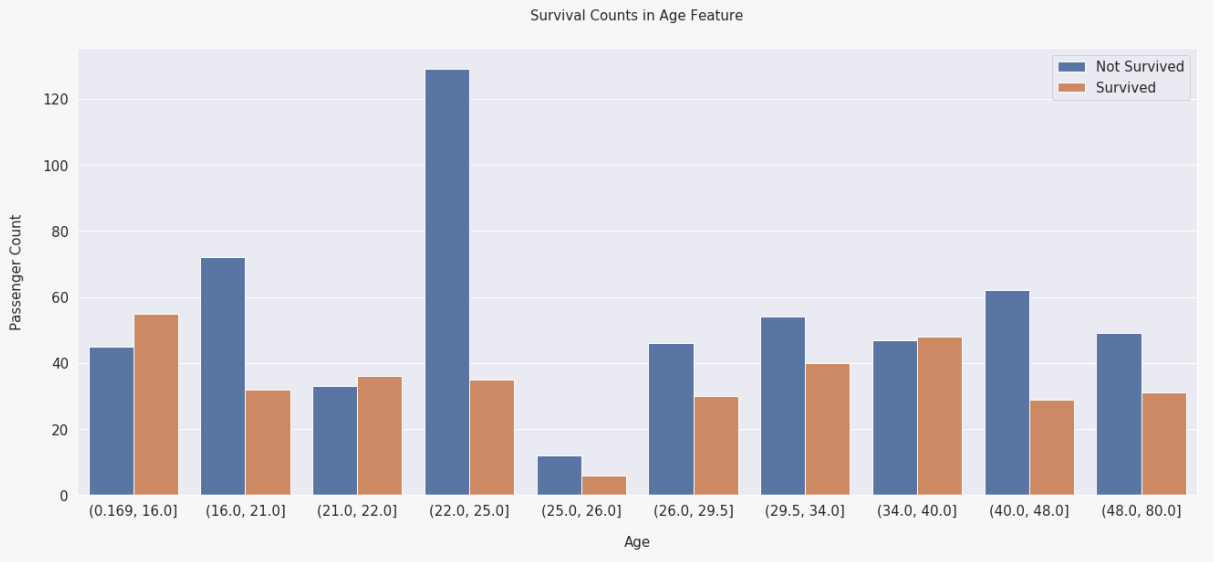


Рисунок 5. Количество выживших, приходящихся на бины Age

Добавим признак Family_Size, создаваемый путем сложения SibSp, Parch и 1. SibSp — это количество братьев, сестер и супругов, а Parch — это количество родителей и детей. Добавление единицы - это учет текущего пассажира. Графики ясно показали, что размер семьи является предиктором выживаемости, потому что разные значения имеют разную выживаемость.

- Размер семьи с 1 обозначить, как как Alone
- Размер семьи с 2, 3 и 4 обозначить, как как Small
- Размер семьи с 5 и 6 обозначить, как как как Medium
- Размер семьи с 7, 8 и 11 обозначить, как как  Large

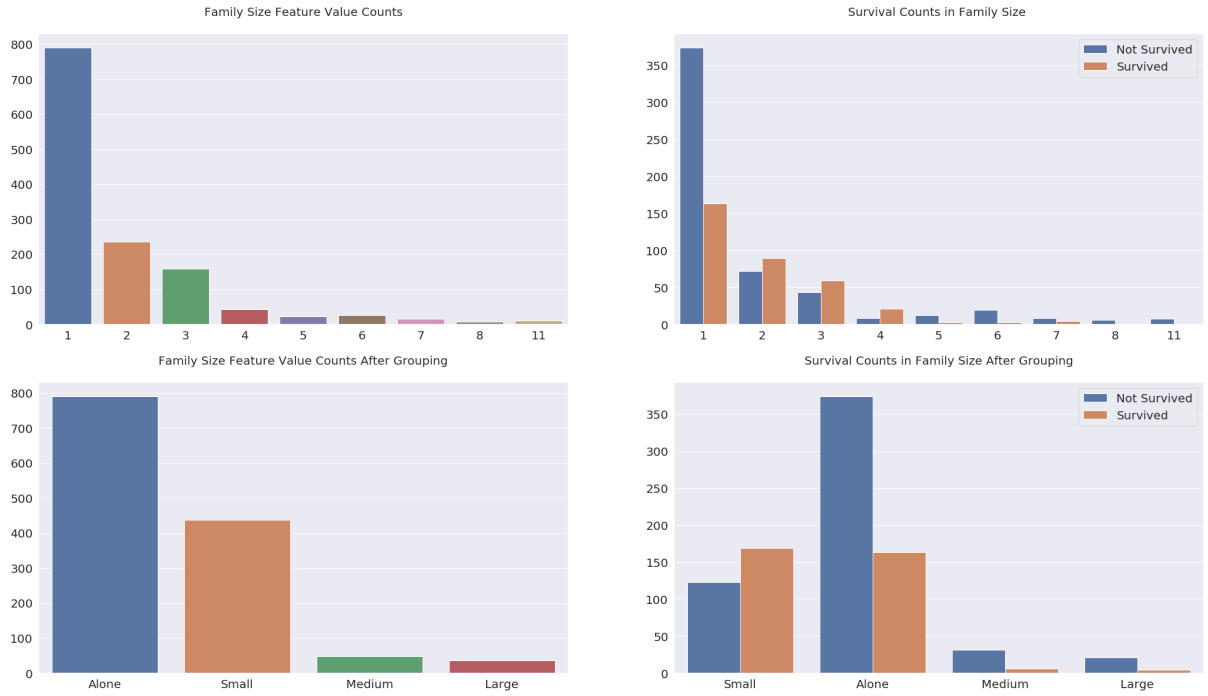


Рисунок 6. Анализ признака Family size

Признак Ticket содержит слишком много уникальных значений для анализа, поэтому его группировка по частоте может упростить задачу.

Чем эта функция отличается от Family_Size? Многие пассажиры путешествовали вместе с группами. Эти группы состоят из друзей, нянь, горничных и т. д. Они не считались семьей, но использовали один и тот же билет.

Если префиксы в признаке Ticket имеют какое-либо значение, то они уже зафиксированы в признаках Pclass или Embarked, потому что это может быть единственная логическая информация, которую можно получить из признака Ticket.

Согласно графику ниже (рис. 7), группы из 2, 3 и 4 пассажиров имели большую выживаемость. Пассажиры, которые путешествуют в одиночку, имеют самый низкий уровень выживаемости. После 4 членов группы выживаемость резко снижается. Этот паттерн очень похож на признак Family_Size, но есть небольшие отличия. Значения Ticket_Frequency не группируются, как Family_Size, потому что это в создает тот же признак с идеальной корреляцией, и такой признак не обеспечивает никакого дополнительного прироста информации.

```
df_all['Ticket_Frequency'] = #Добавьте свой код сюда
```

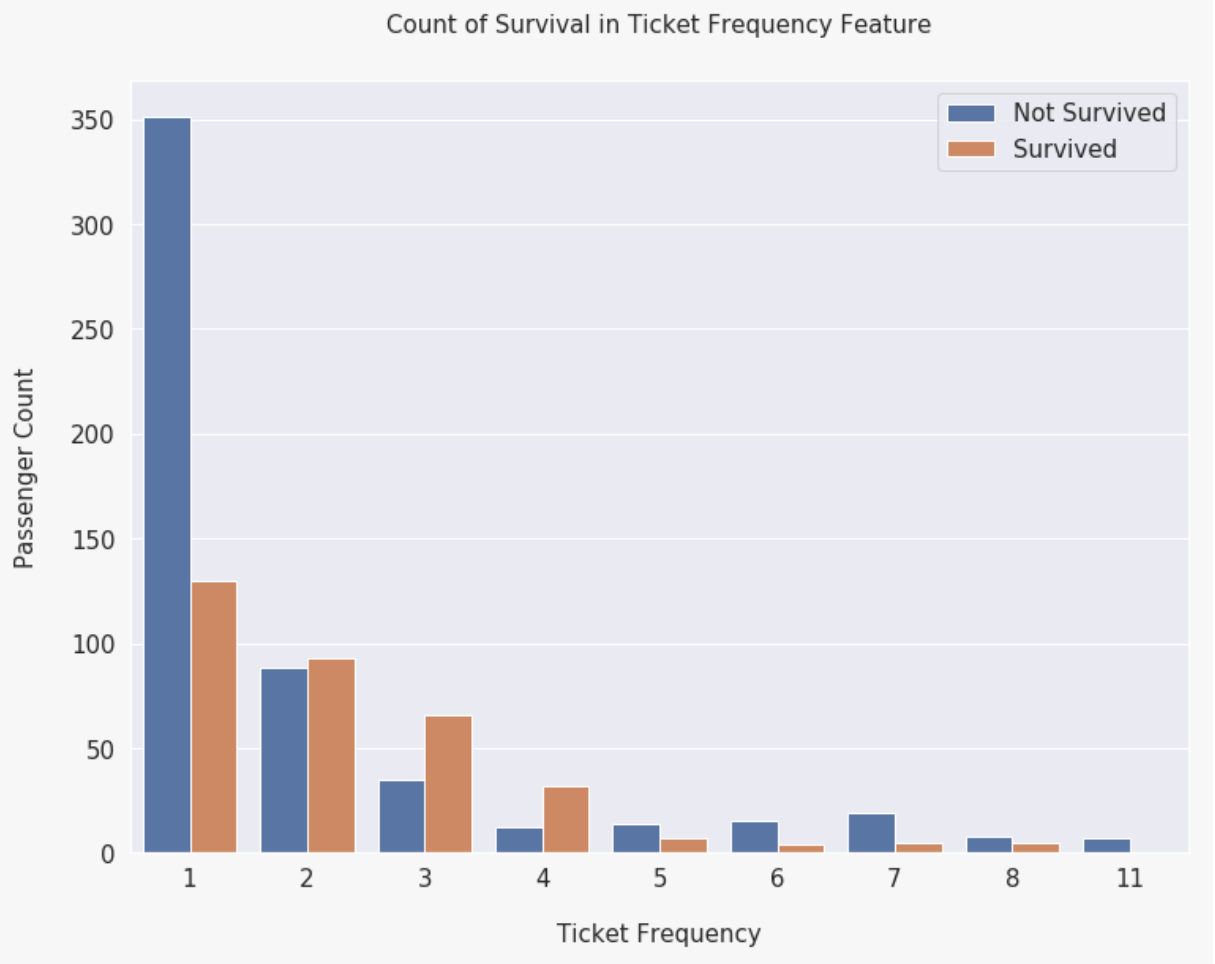


Рисунок 7. Количество выживших пассажиров относительно признака 'Ticket_Frequency '

Title создается путем извлечения префикса в признаке Name. Согласно приведенному ниже графику (Рисунок 8), многие признаки Title встречаются очень редко. Некоторые из них кажутся неправильными, и их необходимо заменить. Признаки Miss, Mrs, Ms, Mlle, Lady, Mme, the Countess, Dona можно заменить на Miss/Mrs/Ms, потому что все относятся к женскому полу. Такие значения, как Mlle, Mme и Dona, на самом деле являются именами пассажиров, но они классифицируются как Title , поскольку функция Name разделена запятой. Признаки Dr, Col, Major, Jonkheer, Capt, Sir, Don и Rev следует заменить на заменены на Dr/Military/Noble/Clergy, потому что эти пассажиры имеют схожие характеристики. Master — уникальное звание, он встречается у пассажиров мужчин моложе 26 лет. У них самая высокая выживаемость среди всех мужчин.

Is_Married — это бинарный признак, основанный на признаке Title Mrs. Признак Mrs имеет самый высокий показатель выживаемости среди других женщин. Признак может выступать особенностью, потому что все женские признаки сгруппированы друг с другом (Miss/Mrs/Ms).

```
df_all['Title'] = #Добавьте свой код сюда
df_all['Is_Married'] = 0
df_all['Is_Married']. #Добавьте свой код сюда
```

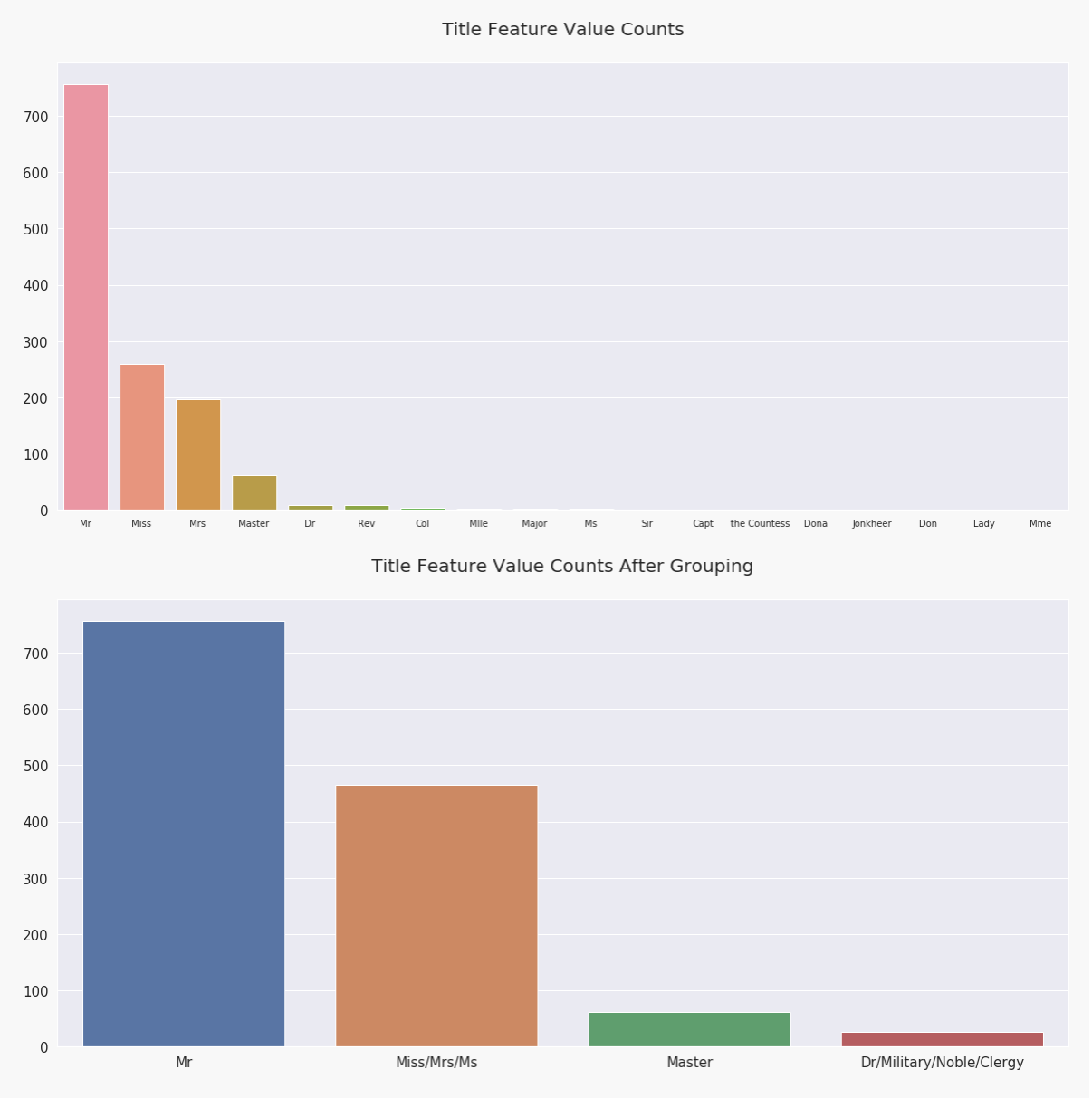


Рисунок 8.

**Задание для выполнения лабораторной работы**

- На основе информации о каюте пассажира сформируйте признак “Палуба” ('Deck') и сгруппируйте значения в нем;
- Разбейте на бины признаки `Fare` и `Age`;
- Вычислите признаки `Family_Size` и `Ticket_Frequency`.
- Добавьте признаки `Title и Is_Married`
- Обучите модель случайного леса на новых данных, полученных в результате преобразований
- Получите значения `Feature Importance`.
- Какие из признаков можно исключить, на основание этих данных?
- Повлияло ли добавление новых признаков на качество предсказаний по сравнению с результатами первой работы?