Titanic Survival Prediction

Robert Mangan

June 19, 2020

Introduction

- Dataset on Titanic passengers including information like age, gender, ticket class, etc., and labels indicating survival status.

Introduction

- Dataset on Titanic passengers including information like age, gender, ticket class, etc., and labels indicating survival status.
- Task: build a model that predicts whether a passenger survived the disaster or not.

Introduction

- Dataset on Titanic passengers including information like age, gender, ticket class, etc., and labels indicating survival status.
- Task: build a model that predicts whether a passenger survived the disaster or not.
- Approach: data cleaning, feature engineering and selection, building and optimising a machine learning model.

Introduction

- Dataset on Titanic passengers including information like age, gender, ticket class, etc., and labels indicating survival status.
- Task: build a model that predicts whether a passenger survived the disaster or not.
- Approach: data cleaning, feature engineering and selection, building and optimising a machine learning model.
- Tools: Python, pandas, scikit-learn.

Dataset

- Training set ($\sim$900 passengers) and test set ($\sim$400 passengers).

Dataset

- Training set (∼900 passengers) and test set (∼400 passengers).

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Dataset

- Training set (∼900 passengers) and test set (∼400 passengers).

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

- Some missing values and data that isn't useful in its current form.

Feature engineering

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

- Extract title (Mr., Mrs., Ms., Dr., etc.) from name.

Feature engineering

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

- Extract title (Mr., Mrs., Ms., Dr., etc.) from name.
- Separate 'SibSp' (sibling + spouse count) into sibling and spouse columns.

Feature engineering

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

- Extract title (Mr., Mrs., Ms., Dr., etc.) from name.
- Separate 'SibSp' (sibling + spouse count) into sibling and spouse columns.
- Count number of passengers sharing a given ticket; calculate fare per person.

Feature engineering

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

- Extract title (Mr., Mrs., Ms., Dr., etc.) from name.
- Separate 'SibSp' (sibling + spouse count) into sibling and spouse columns.
- Count number of passengers sharing a given ticket; calculate fare per person.
- Separate 'Cabin' into 'Deck' and 'Room number'.

Data preparation

- Group by sex/class/title and fill missing age data based on the median value of each group.

Data preparation

- Group by sex/class/title and fill missing age data based on the median value of each group.
- Group by class and fill missing fare data based on the median value of each group.

Data preparation

- Group by sex/class/title and fill missing age data based on the median value of each group.
- Group by class and fill missing fare data based on the median value of each group.
- Fill missing 'embarked' values based on the most frequent value.

Data preparation

- Group by sex/class/title and fill missing age data based on the median value of each group.
- Group by class and fill missing fare data based on the median value of each group.
- Fill missing 'embarked' values based on the most frequent value.
- One-hot encoding for categorical data (e.g. title).

Modelling

- Used random forest classifier.

Modelling

- Used random forest classifier.
- Found optimal hyperparameters (max depth, number of estimators, etc.) using randomized search with cross-validation.

Modelling

- Used random forest classifier.
- Found optimal hyperparameters (max depth, number of estimators, etc.) using randomized search with cross-validation.
- Feature selection (recursive feature elimination).

Results

- ~80% accuracy on the test set.

Results

- ~80% accuracy on the test set.
- Important features included deck, title, age, sex, class.

Questions

Questions?