COMP1816 - Machine Learning Coursework Report

Forename Surname - Student number
Word Count: No more than 2000 words (edit here)

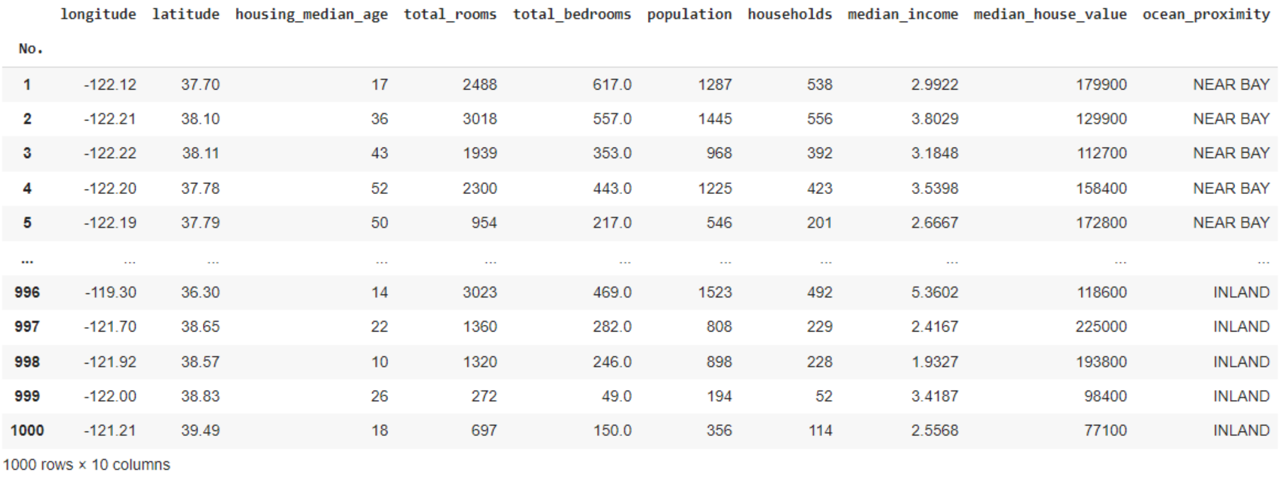
1. Introduction

Machine learning methods are used to evaluate one of the key concepts of real estate analysis called the prediction of housing prices. ML methods also has been implemented in predicting the survival chance of the passengers that had boarded the Titanic. The focus of this project is to build regression model to accurately predict the house median value and a classification model for the prediction of passenger's survival in the Titanic incident. In the regression task, Ridge Regression is the main model implemented which is compared with two baselines models, Linear Regression and Decision Tree Model. In the classification task, Support Vector Machine is the main model which is compared with two baseline models, K Nearest Neighbour (KNN) and Logistic Regression. For the regression task, hyper-parameter tuning is done to get the best results and the results are evaluated using Mean Squared Error (MSE) and R2 score. Ridge Regression model performs the best for the regression task with a R2 score of 0.6653. For classification task, cross-validation is used to obtain the optimal hyper-parameters for the three models and F1 score is considered as the evaluation metric. Support Vector Machine (SVM) performs the best out of three model with a F1 score of 0.7805.

2. Regression

2.1. Pre-processing

The California housing dataset consists of 1000 rows and 10 columns. The 10 columns are divided into 9 features and one target variable called 'median house value' as shown in figure 1.



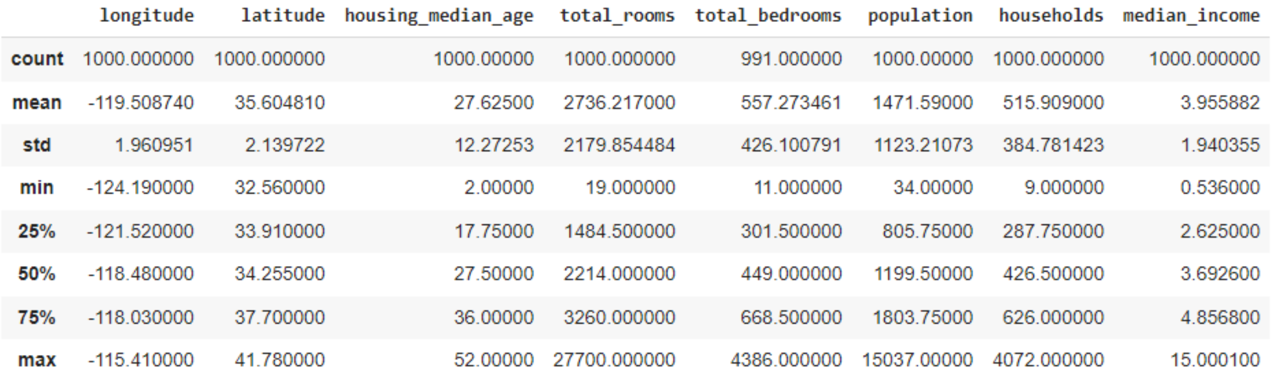
No.	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
1	-122.12	37.70	17	2488	617.0	1287	538	2.9922	179900	NEAR BAY
2	-122.21	38.10	36	3018	557.0	1445	556	3.8029	129900	NEAR BAY
3	-122.22	38.11	43	1939	353.0	968	392	3.1848	112700	NEAR BAY
4	-122.20	37.78	52	2300	443.0	1225	423	3.5398	158400	NEAR BAY
5	-122.19	37.79	50	954	217.0	546	201	2.6667	172800	NEAR BAY
...	...	...	...	...	...	...	...	...	...	...
996	-119.30	36.30	14	3023	469.0	1523	492	5.3602	118600	INLAND
997	-121.70	38.65	22	1360	282.0	808	229	2.4167	225000	INLAND
998	-121.92	38.57	10	1320	246.0	898	228	1.9327	193800	INLAND
999	-122.00	38.83	26	272	49.0	194	52	3.4187	98400	INLAND
1000	-121.21	39.49	18	697	150.0	356	114	2.5568	77100	INLAND

1000 rows × 10 columns

Figure 1. California Housing Dataset

Exploratory Data Analysis(EDA)

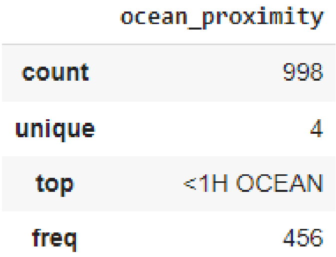
Exploratory Data Analysis is important to better understand the dataset and relation between features and target variables. The dataset is split into 9 features and 1 target variable. The feature column is divided into numerical column and categorical column as as the methods for pre-processing the numerical and categorical data are different. Figure 2 shows that there are 9 missing values in the feature, 'total bedrooms'. The missing values are filled with the median of all the values in that feature.



	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income
count	1000.000000	1000.000000	1000.00000	1000.000000	991.000000	1000.00000	1000.000000	1000.000000
mean	-119.508740	35.604810	27.62500	2736.217000	557.273461	1471.59000	515.909000	3.955882
std	1.960951	2.139722	12.27253	2179.854484	426.100791	1123.21073	384.781423	1.940355
min	-124.190000	32.560000	2.00000	19.000000	11.000000	34.00000	9.000000	0.536000
25%	-121.520000	33.910000	17.75000	1484.500000	301.500000	805.75000	287.750000	2.625000
50%	-118.480000	34.255000	27.50000	2214.000000	449.000000	1199.50000	426.500000	3.692600
75%	-118.030000	37.700000	36.00000	3260.000000	668.500000	1803.75000	626.000000	4.856800
max	-115.410000	41.780000	52.00000	27700.000000	4386.000000	15037.00000	4072.000000	15.000100

Figure 2. California Housing Dataset

The categorical feature called ’ocean proximity’ also consists of 2 missing values as shown in figure 3. These missing values are filled with most repeated category which is ’¡1H OCEAN’ using a simpleImputer.



	ocean_proximity
count	998
unique	4
top	<1H OCEAN
freq	456

Figure 3. California Housing Dataset

Feature Scaling

After dealing with the missing values in the numerical columns, it is important to do feature scaling to the numerical data. Feature scaling is important because it evens out the magnitudes of different features. For example, features like ’median income’ has fractional value of 0.53 while ’total rooms’ contains values that are very large like more than 20,000. There are two common feature scaling methods, Standardisation and Normalisation(MinMaxScaling). From figure 2, it is evident that feature like ’total rooms’ consist of some outliers as the maximum value of that feature is exponentially higher than the rest of the values. Due to its sensitivity to outliers, Normalisation is not recommended for feature scaling in this dataset. Standardisation is considered for feature scaling.

Encoding

For the Machine learning models to use the categorical features, it needs to be converted into numerical values by using a encoder. There are two common encoding methods: Label encoding and one-hot encoding. The categorical feature, ’ocean proximity’ is nominal which means it does not have any order. One-hot encoding is used when the order does not matter. In this case, one-hot encoding is suitable for handling the categorical variable.

This concludes the pre-processing section. Finally, the numerical and categorical columns are combined into a single feature matrix again.

2.2. Methodology

Model selection This paper focuses on the prediction of the median house value with three different regression models, that is Linear Regression Model, Ridge Regression Model and Decision Tree Model. Ridge Regression Model outperforms both polynomial regression model and Decision Tree Model with no signs of overfitting or underfitting.

Ridge regression is a linear regression with the l^2 regularization term. Cost function of ridge regression is slightly different than a linear regression model with the addition of a regularisation function. The cost function of ridge regression is given below:

$$J_\alpha(\theta) = L(\theta) + R_\alpha(\theta) \tag{1}$$

$$L(\theta) = \frac{1}{m}\sum_{i=0}^{m-1}(y^{(i)} - x^{(i)\top}\theta)^2 \tag{2}$$

$$R_\alpha(\theta) = \alpha\frac{1}{m}\sum_{j=1}^{n-1}(\theta_j)^2 \tag{3}$$

Here, L is the loss function that gives the mean squared error R_α is the regularization function that gives l^2 term. α is the regularization weight which is directly proportional to the cost function such that:

- If α is large, regularization is strong and overfitting is strongly avoided but may cause underfitting.
- if α is small, then the regularization is weak which means it might cause overfitting considering there is no change in the model.
- changing the strength of regularization weight(α) is considered as a hyperparameter tuning.

2.3. Experiments

2.3.1. Experimental Settings

Hyperparameter Tuning

All algorithms are run in order to analyze the prediction of 'median house value'. When applying the models to the California housing dataset, it is necessary to tune the hyperparameters in order to get the best results. Instead of changing the strength of hyperparameter of each model manually, a framework is created using for loop and GridSearchCV to automatically increment the hyperparameter with each iteration of fitting the model into training data and making predictions in the train and test data.

Baseline Models

- **Polynomial Regression Model**

  The degree of polynomial is considered as the main hyperparameter of the multivariate regression model. For selecting the best degree of polynomial features for the california housing data,'for loop' is used to initalise the model and increase the hyperparameter variable degree on each iteration. After converting the features to polynomial features, standardisation is applied. This order of preprocessing matters as the purpose of squaring values in PolynomialFeatures is to increase the signal which cannot be possible if the feature scaling is done before the converting to polynomial features as the feature values will get smaller than what it originally was. Then, the Mean Squared Error(MSE) for training and testing data of every iteration is stored in an array. The degree with the lowest MSE in the testing data is considered as the best degree. The model is initialised again with the best degree to get better results on the test data.

- **Decision Tree Model**

Ridge Regression Model

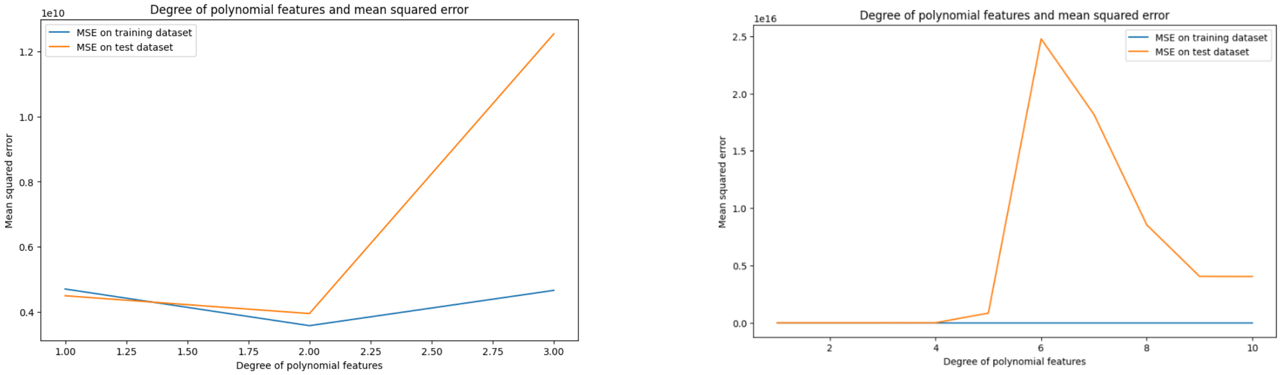
The ridge regression model is similar to baseline regression model which means the polynomial degree is also one of the hyperparameter of ridge regression model. In addition to this, ridge regression model consists of regularization weight (α) which is also a hyperparameter. The same process of automation is applied to this model but $alpha$ value increases with each iteration by keeping the best degree of polynomial.

2.3.2. RESULTS

Baseline Regression Model

Hyperparameter tuning

The Figure 4a shows that MSE for both training and test data is the lowest when the degree of polynomial is 2. In order to find the best degree possible, the model is trained and tested for ten iterations as shown in 4b



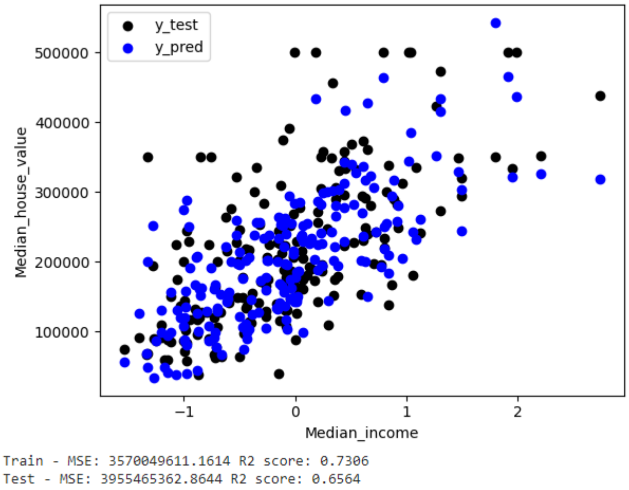
(a) 3 degrees of polynomial features

(b) 10 degrees of polynomial features

Figure 4. Hyperparameter tuning of Multivariate Regression.

Evaluation

MSE and R2 score are the evaluation metrics chosen to analyze the performance of the model. The 5 shows that the MSE value in training data is lower than the MSE value in test data which means that the model resulted in underfitting the data.



```
Train - MSE: 3570049611.1614 R2 score: 0.7306
Test - MSE: 3955465362.8644 R2 score: 0.6564
```

Figure 5. Median income vs Median House value.

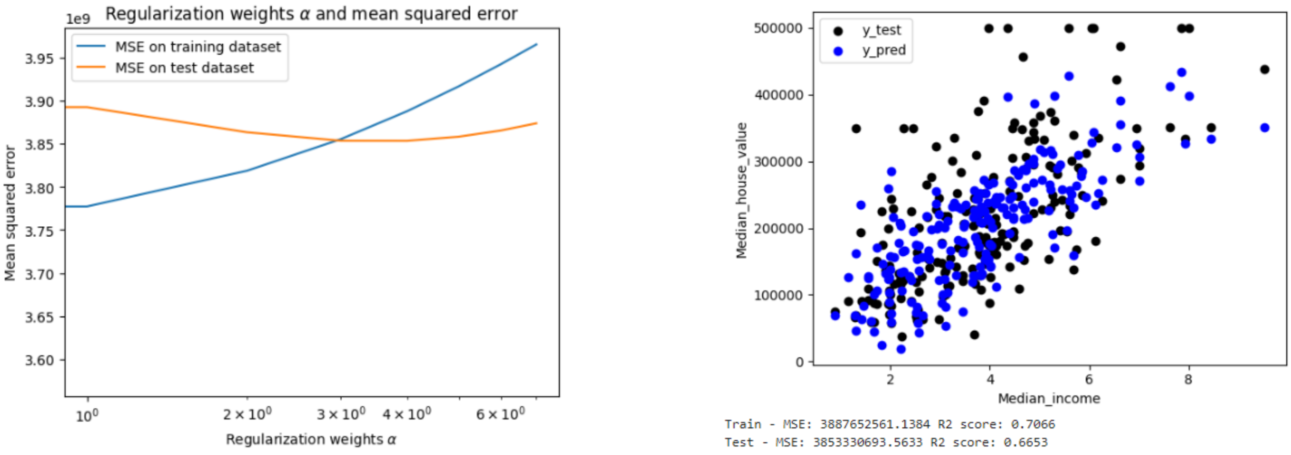
Ridge Regression Model

Hyperparameter Tuning

For ridge regression model, the degree of polynomial is unchanged and is set at 2 as it gave the lowest cost function in the baseline regression model. To further enhance the performance of the model, the regularization weight (α) is tuned to get the lowest MSE without overfitting the training data. The model is fitted into the dataset and predictions are made and the value of α is changed through every iteration of fitting the model to the dataset. The predictions are made in both train and test data and Mean Squared Errors of every iteration are stored in an array. The best α for this model is 3 as there is no sign of overfitting and underfitting as shown in figure 6a.

Evaluation

Since, the feature matrix of multivariate model has many dimensions which cannot be plotted in a 2d graph, one of the feature is selected to visualize the prediction in a graph. In this case, 'Median income' is selected to show the predictions of target feature made by the model and the actual value of the target feature as shown in 6b



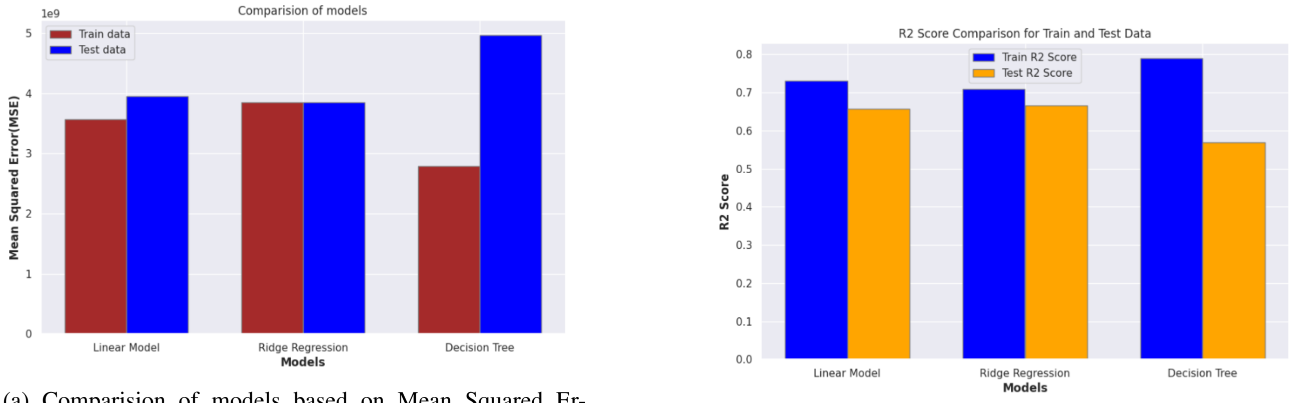
```
Train - MSE: 3887652561.1384 R2 score: 0.7066
Test - MSE: 3853330693.5633 R2 score: 0.6653
```

(a) Hyperparameter Tuning of Ridge Regression Model

(b) Ridge Regression Model

Figure 6. Ridge Regression Model (Tuned)

2.3.3. Discussion



(a) Comparision of models based on Mean Squared Error(MSE)

(b) Comparision of models based on R^2 score

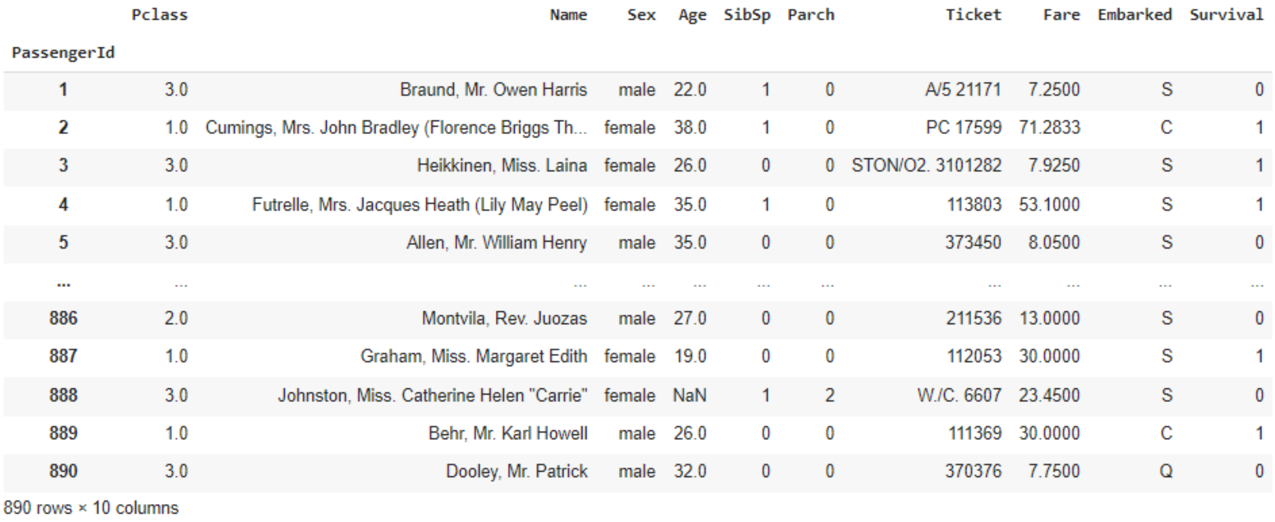
Figure 7. Comparision of regression models

3. Classification

The sinking of Titanic is the most infamous shipwrecks in history. It has been more than a hundred years since this incident took place but research on what impacts the individual's survival rate has still been going on. This paper focuses analyzing the likehood of a passenger's survival and also apply two baseline models: K-Nearest Neighbors(kNN), Logisitic Regression and Support Vector Machine(SVM) to predict the likelihood of survival of a passenger.

3.1. Pre-processing

The Titanic dataset consist of 890 rows and 10 columns with each row representing a individual passenger's attributes. The 10 columns are divided into 9 Feature columns and 1 Target column ('Survival').



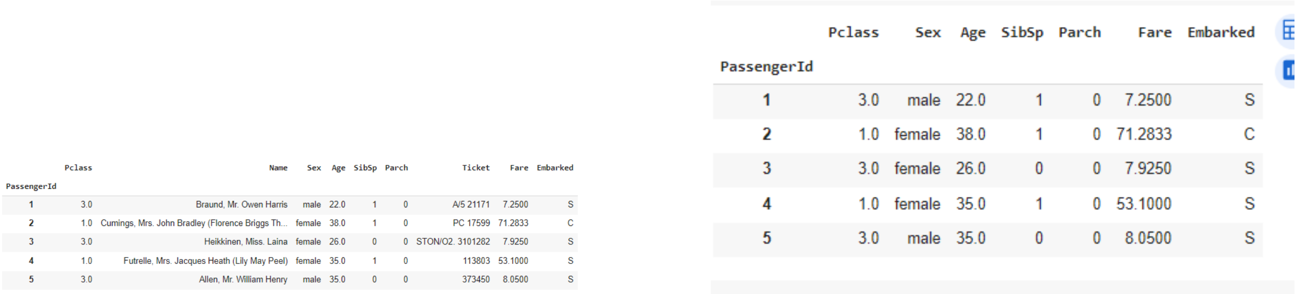
	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked	Survival
PassengerId										
1	3.0	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	S	0
2	1.0	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C	1
3	3.0	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	S	1
4	1.0	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	S	1
5	3.0	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	S	0
...	...	...	...	...	...	...	...	...	...	...
886	2.0	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	S	0
887	1.0	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	S	1
888	3.0	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	S	0
889	1.0	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C	1
890	3.0	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	Q	0

890 rows × 10 columns

Figure 8. Titanic Dataset

EDA

After splitting the dataset to feature columns and target column, Exploratory Data Analysis is performed in the Feature Column/Matrix. There are two features: 'name' and 'Ticket' that has no contribution towards the performance of the model as these are unique identifiers that are unique for each passenger. So, these features can be removed from the feature Matrix as shown in fig 9.



	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
PassengerId									
1	3.0	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	S
2	1.0	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C
3	3.0	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	S
4	1.0	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	S
5	3.0	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	S

(a) Feature Matrix (Before)

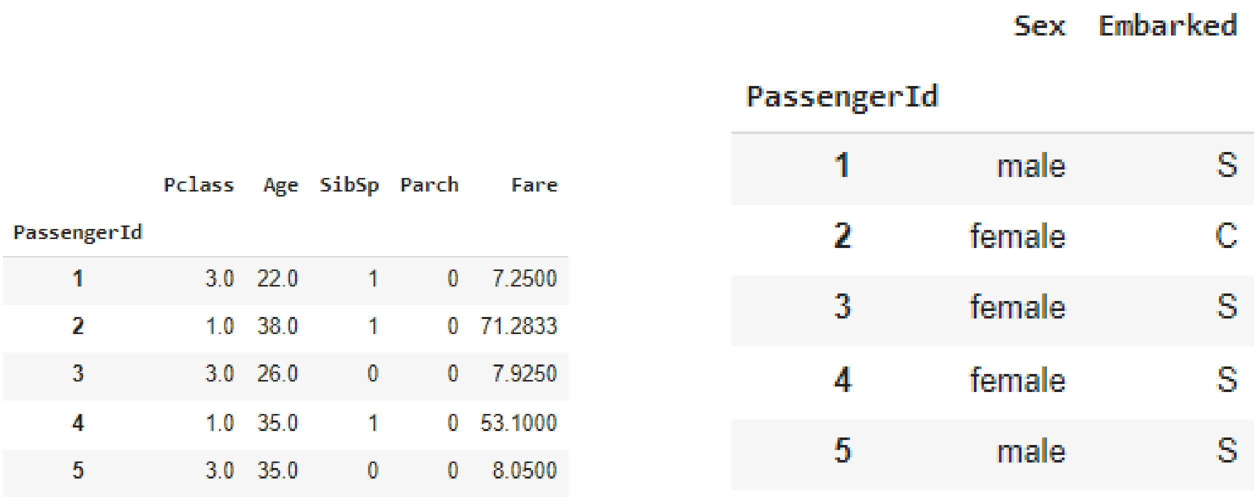
	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
PassengerId							
1	3.0	male	22.0	1	0	7.2500	S
2	1.0	female	38.0	1	0	71.2833	C
3	3.0	female	26.0	0	0	7.9250	S
4	1.0	female	35.0	1	0	53.1000	S
5	3.0	male	35.0	0	0	8.0500	S

(b) Feature Matrix (After)

Figure 9. Feature matrix

Then, the numerical and categorical data in the feature matrix are seperated to preprocess them as shown in Fig 10

A detailed description of the numerical feature columns are shown in fig 11 There are some missing data in the numerical feature 'Age' which is filled using the median value of that feature. Fig 11 shows that the minimum value of 'Fare' is



PassengerId	Pclass	Age	SibSp	Parch	Fare
1	3.0	22.0	1	0	7.2500
2	1.0	38.0	1	0	71.2833
3	3.0	26.0	0	0	7.9250
4	1.0	35.0	1	0	53.1000
5	3.0	35.0	0	0	8.0500

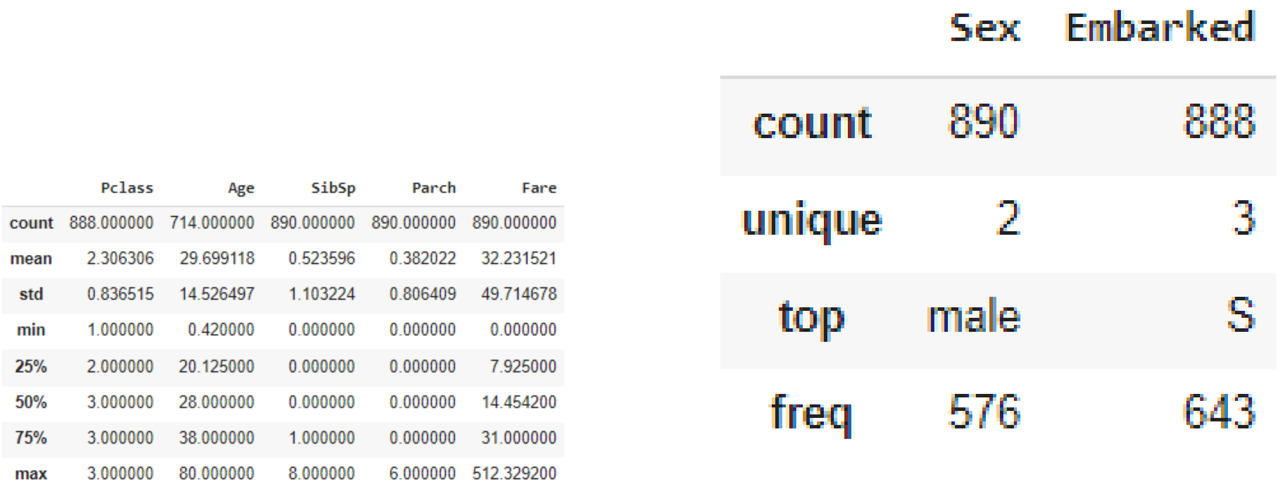
(a) Numerical Features

PassengerId	Sex	Embarked
1	male	S
2	female	C
3	female	S
4	female	S
5	male	S

(b) Categorical Features

Figure 10. Data Splitting

zero. When identifying the passengers who paid no fare at all, some 1st class passenger were present which is not possible considering 1st class passengers have been charged a higher amount of fare, and it is obvious that every passenger has paid the fare. This issue is solved by converting the zero valued fare into NaN values which is later impuded using same method that was used in the 'Age' feature.



	Pclass	Age	SibSp	Parch	Fare
count	888.000000	714.000000	890.000000	890.000000	890.000000
mean	2.306306	29.699118	0.523596	0.382022	32.231521
std	0.836515	14.526497	1.103224	0.806409	49.714678
min	1.000000	0.420000	0.000000	0.000000	0.000000
25%	2.000000	20.125000	0.000000	0.000000	7.925000
50%	3.000000	28.000000	0.000000	0.000000	14.454200
75%	3.000000	38.000000	1.000000	0.000000	31.000000
max	3.000000	80.000000	8.000000	6.000000	512.329200

(a) Summary of Numerical Columns

	Sex	Embarked
count	890	888
unique	2	3
top	male	S
freq	576	643

(b) Summary of Categorical Columns

Figure 11. Detailed Description of Feature Matrix

3.2. Methodology

Support Vector Machine

Support Vector Machines are another very popular machine learning method (which can be used for regression and classification). The mathematics behind them are quite complicated and different to what we have been looking at in our linear models - using objects called 'kernels' to draw the line that best separates the classes, rather than regressing on the probability like our Logistic Regression (as such, we don't generally get a probability estimate from the SVM).

There are a few key settings you should consider for SVMs:

c - The inverse Regularisation Weight (there are no options of the type of regularisation for SVM).

kernel - the type of kernel the SVM model uses to fit the lines

linear - Draws a straight line

poly - Same as manually fitting polynomial features and using the linear kernel (though doing it this manual way would be computationally slower).

degree - Choose maximum polynomial degree

rbf - Uses Kernel trick to measure similarity between data points in infinite dimensions and classify based on this.

sigmoid - Creates a complex sigmoid shaped irregular boundary. Only generally useful for very specific datasets as is often does not generalise well.

random state - Allows you to recreate the same solution by generating the same quasi-random case.

3.3. Experiments

3.3.1. Experimental Settings

Baseline Models

3.3.2. Results

Logistic Regression

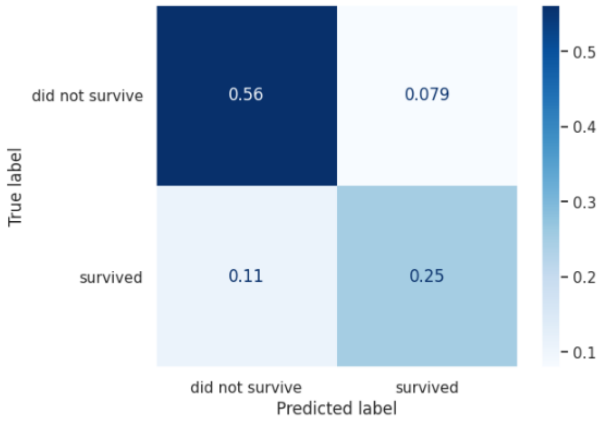


Figure 12. Confusion Matrix

k-Nearest Neighbors

Support Vector Machine

3.3.3. Discussion

4. Conclusion

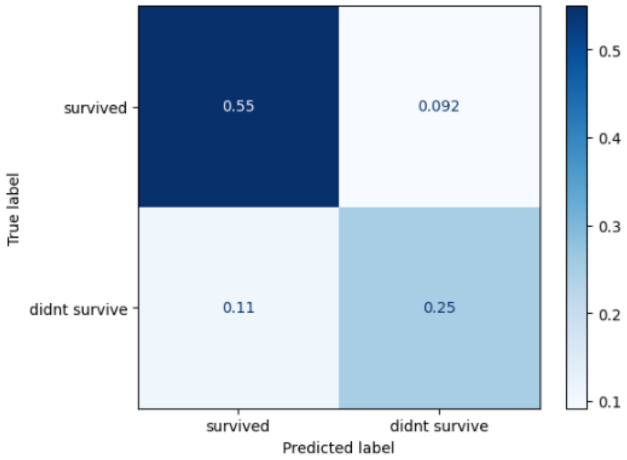


Figure 13. Confusion Matrix

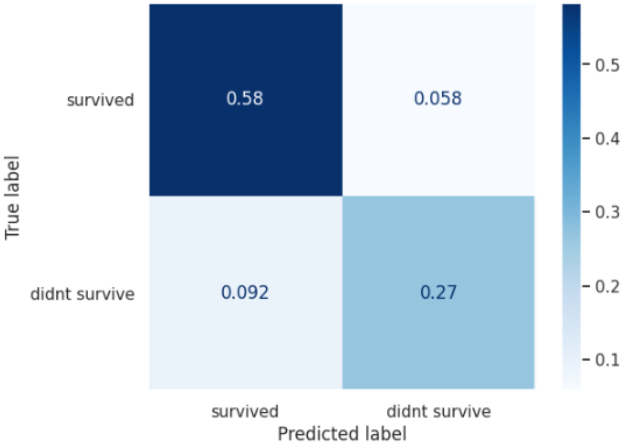


Figure 14. Confusion Matrix