Executive Summary/Abstract

In this project, the posterior estimates using MCMC for the coefficient of 7 features are obtained. We engineered new feature which improved accuracy and after training, convergence was observed. This is verified using the trace plot, autocorrelation and other posterior checks. The model has a 79 percent accuracy on out of sample data.

Introduction

For the purpose of the project, the titanic dataset would be used. The dataset is aimed at predicting if a passenger, given some features, aboard the titanic died or not. The data can be downloaded from Kaggle website via the link: https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/stuff/titanic.csv. The dataset consists of both training and testing set. The training set has 891 rows while the test set has 418 rows. There are 8 features observed including the target variable. This project would make use of the training set for both training and validation. The first 2 row of the dataset is shown in Table 1:

Table 1: Table showing the first two row of the dataset

Survived	Pclass	Name	Sex	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare
0	3	Mr. Owen Harris Braund	male	22.0	1	0	7.2500
1	1	Mrs. John Bradley (Florence Briggs Thayer) Cum...	female	38.0	1	0	71.2833

Data

The 8 features provided for each datapoint are, Survived, Pclass, Name, Sex, Age, Siblings/Spouses Aboard, Parents/Children Aboard, fare. The response variable is Survived. The statistics of each feature is shown in Table 2:

Table 2: Table showing the statistics of the numeric variable

	Survived	Pclass	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare
count	887.000000	887.000000	887.000000	887.000000	887.000000	887.00000
mean	0.385569	2.305524	29.471443	0.525366	0.383315	32.30542
std	0.487004	0.836662	14.121908	1.104669	0.807466	49.78204
min	0.000000	1.000000	0.420000	0.000000	0.000000	0.00000
25%	0.000000	2.000000	20.250000	0.000000	0.000000	7.92500
50%	0.000000	3.000000	28.000000	0.000000	0.000000	14.45420
75%	1.000000	3.000000	38.000000	1.000000	0.000000	31.13750
max	1.000000	3.000000	80.000000	8.000000	6.000000	512.32920

The categorical variables provided are sex and Name. Sex can either be male or female, and names are unique identifier for each row. Both features are categorically encoded, and their statistics are shown in Table 3:

Table 3: Table showing the statistics of categorical variable

Statistics	Name	Age
Count	887	887
Mean	0.354002	443.000000
Std	0.478480	256.199141
Min	0.00	0.00
25%	0.00	221.5
50%	0.00	443.0
75%	1.00	664.5
100%	1.00	886.0

In terms of null variable, there are no null values and thus, all rows can be used in training the model. However, name columns would be dropped as it provides no information since all values are unique as observed from table 3.
The pair plot showing the interaction between the covariates and response variable is shown in Figure 1.

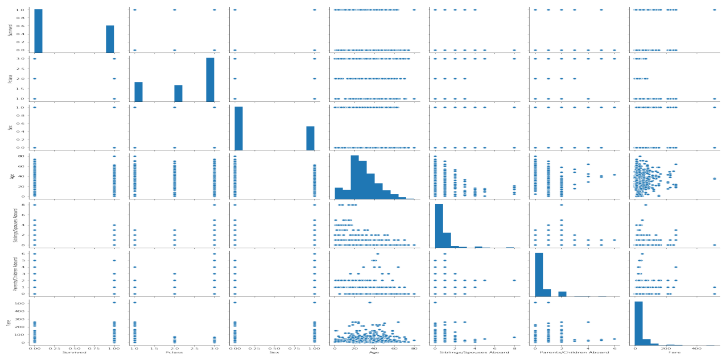

Figure 1: Pair Plot showing the interaction between the covariates and response variable

From the plot, siblings aboard, parent children and fare have their histogram skewed to the right. Also, from figure 2, we can see a negative correlation between both siblings abroad and parent/children abroad and the response variable. The survival chance decreases with increase in Parents/ Children Aboard and Siblings/Spouse Aboard

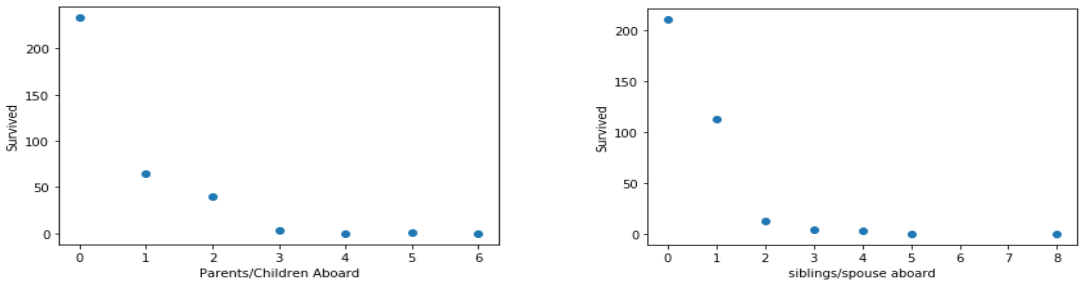

Figure 2a & b: The plots show the interaction between Survival and Parents/Children Aboard and Sibling/Spouse Aboard respectively

The correlation between features is shown in figure below

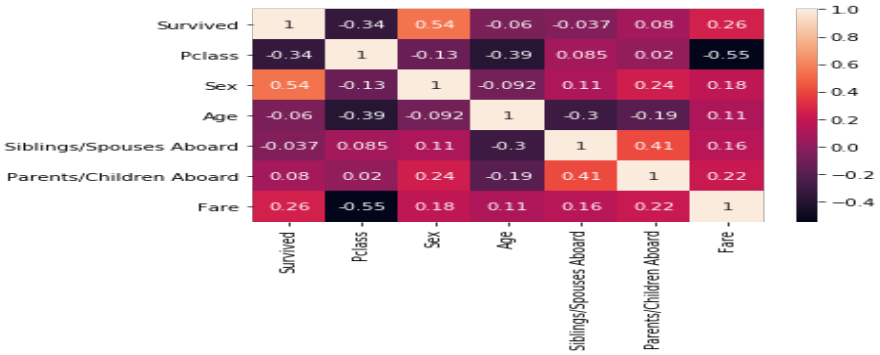

Figure 3: Autocorrelation plot for the features

Model

The age and fare features are normalized to mean 0 and variance one for modelling. We shuffled our dataset and used the first 700 observation to train the model, and left out the remaining observation to validate the result. Bernoulli distribution is used to obtain the maximum likelihood estimate since this is a logistic regression task. The coefficients have a Normal prior with mean 0.0 and non-informative variance of 100^2. We created a new feature using Age and pclass features by string concatenating and categorical encoding each observation. A burn in of 1000 is allowed, and a draw of 4000 for 3 different chains is used.

Results

From the trace plot, the mcmc trace plot is observed to have converged for the coefficients. This is also shown using the Gelman rubin and autocorrelation plot. The gelman rubin for all features is close to 1.00 indicating convergence as shown in table 4.

Table 4: Table showing the result of Gelman Rubin evaluation on the trained model

Coefficient of Data variables	Gelman Rubin
Intercept	1.001
b_class	1.001
b_sex	1.0
b_class_sex	1.001
b_sibling	1.0
b_child	1.001
b_fare	1.0
b_age	1.0

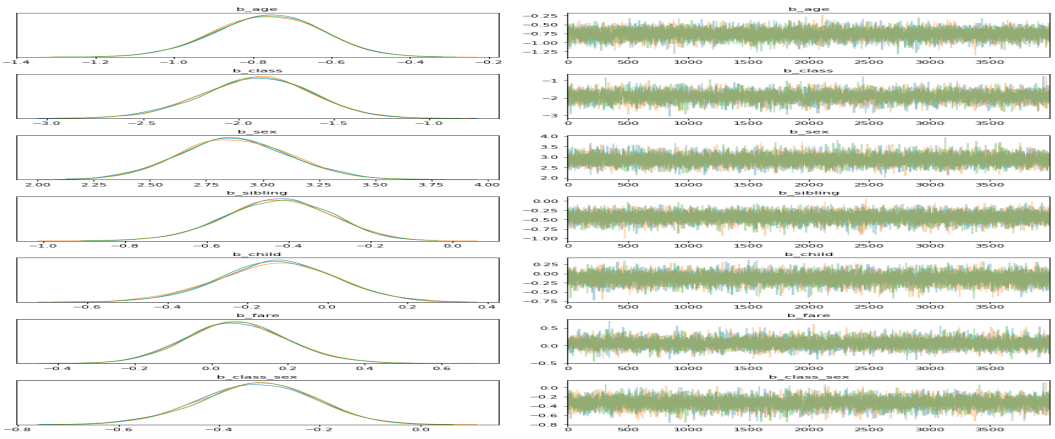

Figure 4: Trace Plot for the trained models

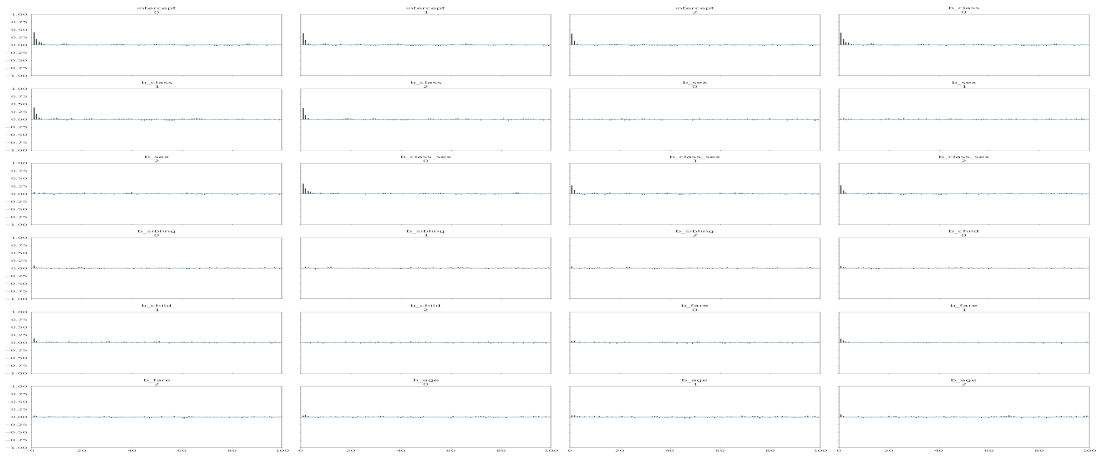

Figure 5: Autocorrelation Plot for different chains of 4000 draws

From obtaining the deviance information criteria, we obtained an effective parameter of 8.12, the DIC of 601.36 using the posterior result.

When evaluated on the test set, the model obtained an accuracy of 79% and a train accuracy of 81%. The confusion matrix is shown below

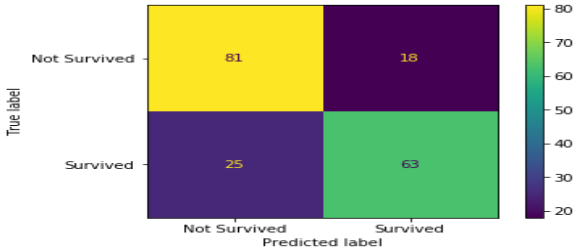

Figure 6: Confusion matrix of the validation set

The posterior estimate of the coefficients is shown in table 5.

Table 5: Posterior Estimate of the coefficients

	mean	sd	hpd_3%	hpd_97%	mcse_m	mcse_s	ess_me	ess_sd	ess_bul	ess_tai	r_ha
intercept	3.546	0.825	2.076	5.215	0.012	0.008	4971	4711	5019	5766	1
b_class	-1.909	0.283	-2.443	-1.382	0.004	0.003	4956	4772	5023	6236	1
b_sex	2.874	0.242	2.436	3.347	0.002	0.002	11328	11328	11299	8097	1
b_class_sex	-0.327	0.114	-0.547	-0.118	0.002	0.001	5626	5574	5655	7037	1
b_sibling	-0.428	0.128	-0.672	-0.194	0.001	0.001	10453	9711	10555	8196	1
b_child	-0.127	0.146	-0.4	0.151	0.001	0.001	9873	7052	9944	8082	1
b_fare	0.066	0.131	-0.183	0.311	0.001	0.001	8966	5476	9096	6833	1
b_age	-0.763	0.135	-1.016	-0.509	0.001	0.001	9602	9376	9637	7879	1

Conclusions

In this project, we were able to sample multiple draws from the posterior estimate of the coefficient of the chosen covariate. The trace plot showed convergence and our model is observed to perform well on out of sample data. Further works would involve engineering new features, using ANOVA for the categorical variables and using a different prior like t-student or Laplace distribution as the prior for some coefficients.