



Exploratory Data Analysis and Machine Learning on Titanic Disaster Dataset

A Report Submitted
in Partial Fulfillment of the Requirements
for the Degree of
Bachelor of Technology
in
Information Technology

by
Abhinav (20168013), Aakarsh Verma (20168002),
Abhishek Dixit (20168004) and Manoj Mahour (20158048)
Group: IT-21

to the
COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
MOTILAL NEHRU NATIONAL INSTITUTE OF TECHNOLOGY
ALLAHABAD

May, 2020

UNDERTAKING

I declare that the work presented in this report titled “*Exploratory Data Analysis and Machine Learning on Titanic Disaster Dataset*”, submitted to the Computer Science and Engineering Department, Motilal Nehru National Institute of Technology, Allahabad, for the award of the ***Bachelor of Technology*** degree in ***Information Technology***, is my original work. I have not plagiarized or submitted the same work for the award of any other degree. In case this undertaking is found incorrect, I accept that my degree may be unconditionally withdrawn.

May, 2020
Allahabad

(Abhinav 20168013)

(Aakarsh Verma 20168002)

(Abhishek Dixit 20168004)

(Manoj Mahour 20158048)

CERTIFICATE

Certified that the work contained in the report titled “*Exploratory Data Analysis and Machine Learning on Titanic Disaster Dataset*”, by *Abhinav, Aakarsh Verma, Abhishek Dixit and Manoj Mahour*, has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

(Er. Manoj Wariya)

Computer Science and Engineering Dept.
M.N.N.I.T. Allahabad

May, 2020

Preface

The Machine Learning has seen an enormous growth in the recent past. With the systems getting smarter and automated, the days are not far when we will become completely dependent on them. The most important feature for success of a system is the amount of data that went into training it and the amount of data to which it gives correct results. Thus, data plays the most important role in Machine Learning models, as it is the result of the data available that will decide the future of the system at hand.

The data is present in enormous forms and formats. Thus, it becomes important to study the kind of data and select an algorithm that works the best with that kind of data.

Acknowledgments

After an intensive period of learning and developing, this note of acknowledgement is our final touch on our report. We would like to express our deep gratitude and sincere thanks to everybody who has helped us in completing the project. First and foremost, we would like to thank our mentor and supervisor, Er. Manoj Wariya, for giving us this opportunity and providing constant support, guidance and encouragement. His innovative ideas and zeal to motivate and help us have led to the successful completion of this project. He has provided us ample opportunity to explore the content and dimensions of this project. We felt privileged working under him.

We would also like to express our sincere gratitude to Prof. Rajeev Tripathi, Director, MNNIT Allahabad, Prayagraj, Prof. Anil Kumar Singh, Head, Computer Science And Engineering Department and Dr. Shashwati Banerjea, Head, Departmental Under Graduate Committee (DUGC) for providing us with the tools and facilities to complete the project.

Finally, we would like to thank our friends and family for their constant support and advice. Without their love, blessings and encouragement it would have proved impossible to complete this project.

Abstract

The Royal Mail Steamer, TITANIC, was the largest cruise ship ever made. The British cruise ship collided during its only journey, with a huge iceberg. The collision happened in the Pacific Ocean, when the Cruise was moving from Southampton to the New York city. There were approximately 2400 passengers on the Cruise when the accident happened and more than half of them could not survive. This unfortunate yet one of the biggest incident forces the researchers and data analysts to analyse and go deep in the data set. The aim of the various studies going on is to explore the data available and find a pattern and impact of various features on the survival of a person if he/she was on the ship.

The survival of the passengers has been analysed using various different algorithms and they have been compared. A new algorithm has been proposed that will give more accurate results than all the previously analysed algorithms, using the features of those algorithms only.

Contents

Preface	iv
Acknowledgments	v
Abstract	vi
1 Introduction	1
1.1 Motivation	2
2 Related Work	3
3 Proposed Work	4
3.1 Objective	4
3.2 Software Requirement Specification	5
3.2.1 Introduction	5
3.2.2 Requirement Specification	6
3.3 Dataset	7
3.4 Building Machine Learning Models	8
3.4.1 Logistic Regression	8
3.4.2 K-Nearest Neighbours	10
3.4.3 Decision Tree	12
3.4.4 Random Forest	14
3.4.5 Decision Tree Hypertuning	16
3.4.6 Support Vector Machines	18
3.4.7 Stochastic Gradient Descent	20

3.4.8	Perceptron	22
3.4.9	Naive Bayes	24
3.4.10	Stacking	26
4	Results And Analysis	28
4.1	Results	28
4.2	Analysis	30
5	Conclusion and Future Work	32
5.1	Conclusion	32
5.2	Applications	33
5.3	Future Work	33
	References	34

Chapter 1

Introduction

The Machine Learning has acquired an inevitable position in today's world. Everything is getting automated and we rely on some software to predict values for us, analysing which, we make the important decisions. These software need to be trained using enormous amount of data so that they understand the underlying pattern and develop a knowledge based on its observations. Then this knowledge is used to analyse any data and observations are made. These observations are correlated to the previous events of that type which have been used to train the software. Hence, these values are highly accurate and important future decisions can be made keeping those observations in reference.

With the growth of such systems, a lot of researchers are working on huge amounts of data, trying to gather as much useful data as possible and analysing it and training the models using this data which then will be used to predict more values. The quality of data is an important aspect here. The data available is falling short of the demands of the researchers, maybe because sufficient information is not available or there has been no work in that particular field.

In this project, we have analysed the information of the passengers that were present on the RMS Titanic when it collided with the Iceberg. Only 712 out of the 2456 people present could survive the mishap. We have worked upon:

1. The attributes of the people who survived.
2. Does the gender of the person determined the chance of survival.

3. Does the class of the Compartment determined their chances of survival.
4. We have analysed the data available on various models including Decision Tree, Logistic Regression, etc. and compared their results.
5. A new algorithm has been proposed which gives more accurate predictions than the existing models.

Understanding the data is the key for all the analytical processes. The analysis has been done on Jupyter Notebook in Python language.

1.1 Motivation

This is the era of Machine Learning, with everything getting automated and this has even made the process of making decisions depending upon the previous similar incidents automated. The models need to be trained on a number of previous incidents and they are capable of predicting the future based on the recognised pattern.

With the growth of such systems, a lot of researchers are working on huge amounts of data, trying to gather as much useful data as possible and analysing it and training the models using this data which then will be used to predict more values. The quality of data is an important aspect here. The data available is falling short of the demands of the researchers, maybe because sufficient information is not available or there has been no work in that particular field.

This motivated us to analyse the data available of the people on the titanic ship and make observations. The aim is to find the underlying pattern and then to predict chances of survival of a person had he/she been on the ship.

Chapter 2

Related Work

The data analysis done is used for building of both Predictive as well as descriptive models. The predictive model allows us to get the missing values in the data set and then predict if a person would survive in such a scenario or not. This may be used to predict if such incident could reoccur in future. The descriptive model is useful for the first model, as it will tell which features are of more importance while predicting the survival chance.[1]

Ju Liu has provided the information regarding the need of analysing the data in segmented manner and finding the correlations and establishing relations and then predicting using this.[2]

Lin, Kunal and Zeshi have established the fact that it isn't necessary that the model's accuracy will improve if more features are provided to it. It is also stated that the dimensionality reduction plays an important role too.[3]

The data set of the people present on the ship is provided by the Kaggle team.[4]

Chapter 3

Proposed Work

3.1 Objective

The main objective of our project is to analyse the data available of the people who were present on the Titanic when the ship collided with the Iceberg and drowned. We have done exploratory data analysis on the features such as Gender, Age, Class of Compartment, etc. of the passengers using a number of Predictive models. These models predict if the person would have survived or not if he/she was on the ship.

We have then compared all the models and used the results to produce an algorithm that has the best accuracy out of all the existing predicting models. Our work includes the comparison of existing models to determine which gives the best results in such situations and then design an algorithm that can be used to predict the chances of survival to avoid any such loss in future.

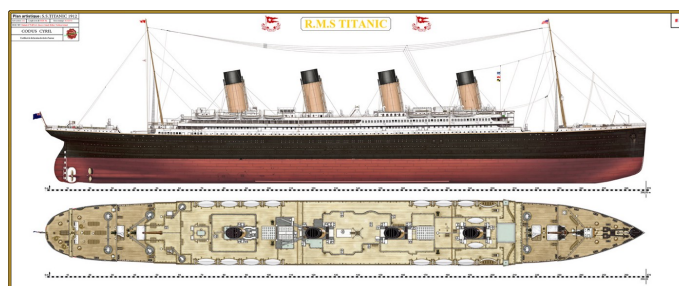


Figure 1: The TITANIC

3.2 Software Requirement Specification

3.2.1 Introduction

■ *Purpose*

The aim of this document is to provide a detailed analysis of the data of the passengers of the RMS Titanic. The various models are compared on the basis of their accuracy to predict if a person survived or not using the information available. In the last, using the results, one algorithm has been proposed that suits the best for such kind of data.

■ *Scope*

The scope of our work extends to both the predictive and the descriptive models. The data that is being considered in this case is of a particular incident, and it is assumed to be correct. Using the analysis done on this data, new model has been proposed that could be used to predict the chances of survival of a person in any of such unforeseen event.

■ *Overview of Document*

The next section, the Overall Description section, of this document gives an overview of the functionality of the project. It describes the informal requirements and is used to establish a context for the technical requirements specification in the next section. The third section, Requirements Specification section, of this document is written primarily for the developers and describes in technical terms the details of the functionality of the product. Both sections of the document describe the same software product in its entirety, but are intended for different audiences and thus use different language.

3.2.2 Requirement Specification

■ *Functional Requirements*

The analysis has been done on Jupyter Notebook using Python Language. The basic requirement is of sufficient data to train the Machine Learning models. The data should consist of a number of attributes, features so that the analysis can be done on a broader scale and more accurately.


■ *Non-Functional Requirements*

The proposed work is more of analysis and hence we will get better results if the data available is vast and true. The quality of data plays a vital role. The accuracy of various models has been compared to improve the performance of the system and a new algorithm has been devised which will further improve the performance and accuracy.

3.3 Dataset

The data has been taken from Kaggle.com, where it is available in the Comma-Separated format. The data set contains 891 rows with attributes including the name of the passenger, the number of siblings, the number of parents or children, the cabin, the ticket number, the fare of the ticket and the place where the person is from.[3]

Pre-processing of the data had to be done because the data had missing values and also some data was present in string format, which had to be converted into numeric types so that our model could consider it for analysis. The missing values have been filled with the median of the available values. The data has been split into training and testing to calculate the efficiency of our models. Before the algorithm for the models is built, a number of exploration graphs have been plotted to find out the features which will influence the model the most.



Attributes	Description
Passenger ID	Identification Number of Passenger
P-class	Passenger Class (1,2,3)
Name	Name of The Passenger
Sex	Gender of The Passenger (Male, Female)
Age	Age of The Passenger
Sib Sp.	Number of Sibling or Spouse on The Ship
Parch	Number of Children or Parent on The Ship
Ticket	Ticket Number
Fare	Price of The Ticket
Cabin	Cabin number of the Passenger
Embarked	Port of Embarkation
Survived	Target Variable (Value 0 For Perished, 1 For Survived)




Figure 2: Data Attributes

3.4 Building Machine Learning Models

We have trained several Machine Learning models and compared their results. The data set did not provide labels for the training-set, hence we use the predictions on the training set to compare the models. Later, we have compared using cross validation as well.

3.4.1 Logistic Regression

1. This statistical algorithm is used to predict the probability of binary outcomes based on one or more independent variables. It means, this is used to predict an outcome which has 1 or 0 ,yes or no, pass or fail.
2. Probabilities are estimated using sigmoid function. The graph of sigmoid function is as:

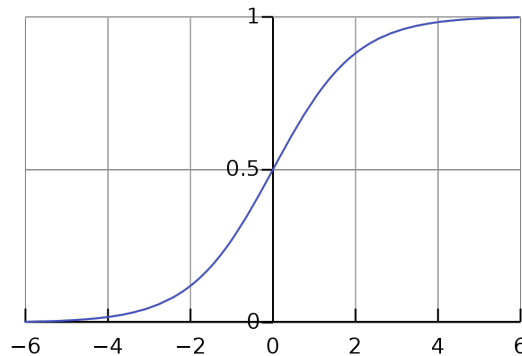


Figure 3: Sigmoid Function

3. Mathematical expression of the sigmoid function is given by:

$$F(x) = 1/(1 - e^{-z})$$

Where $z=w_0+w_1*x_1+w_2*x_2+.....+w_n*x_n$ and $x_1,x_2,x_3,...,x_n$ are independent variables and $F(x)$ is probability of binary outcome.

4. Values greater than 0.5 have been classified as survived (1) and others to 0.
5. We get a score of **81.36% with 214 correct predictions.**

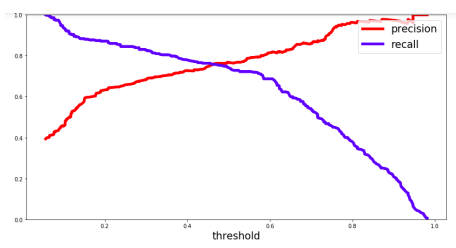


Figure 4: Recall, Precision vs Threshold for Logistic Regression

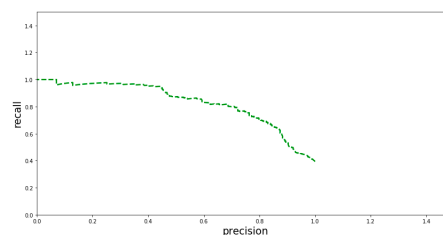


Figure 5: Recall vs Precision for Logistic Regression

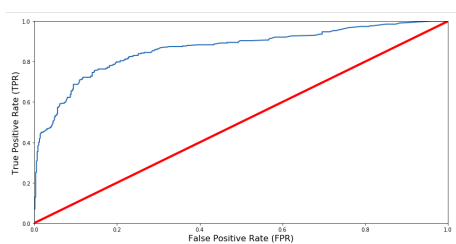


Figure 6: True Positive vs False Positive for Logistic Regression

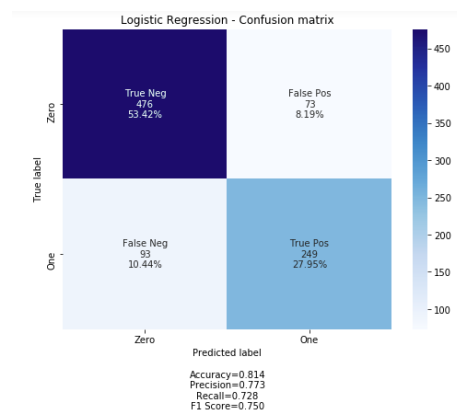


Figure 7: Confusion Matrix for Logistic Regression

3.4.2 K-Nearest Neighbours

1. K nearest neighbours is a classification algorithm highly useful in classifying an unknown data set primarily on the similarity of the neighbouring results.
2. The object is assigned its membership to a class by majority vote of its k nearest neighbours.
3. The degree of 'closeness' is calculated by distance of a said 'object' to its k neighbours.

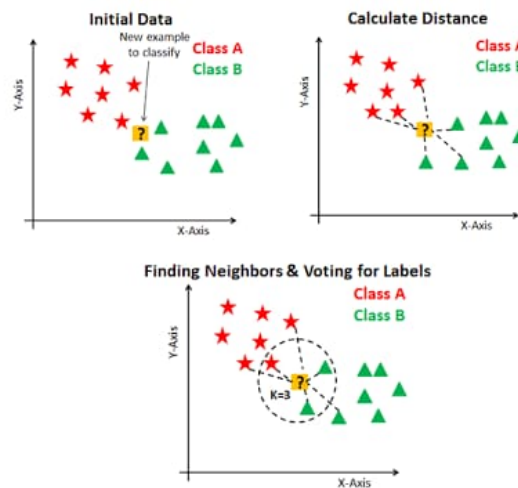


Figure 8: Working of KNN algorithm

4. The model is trained with the value of k equal to 3 which yielded **81.03%** with **232 correct** predictions.

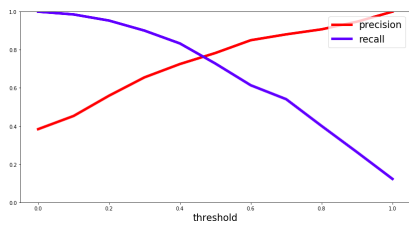


Figure 9: Recall, Precision vs Threshold for KNN

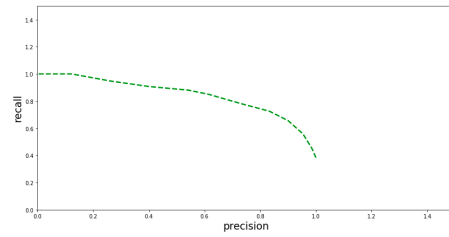


Figure 10: Recall vs Precision for KNN

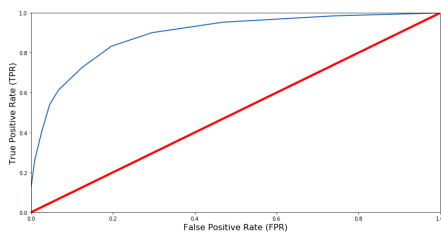


Figure 11: True Positive vs False Positive for KNN

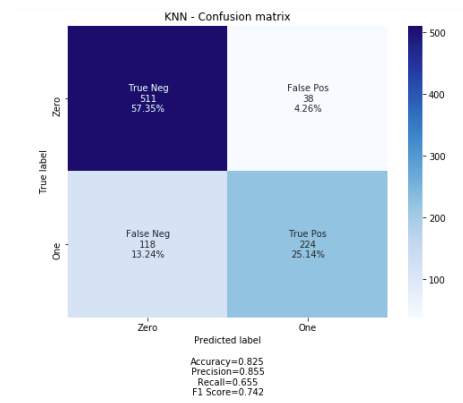


Figure 12: Confusion Matrix for KNN

3.4.3 Decision Tree

1. In decision tree, the results are branched according to a particular condition and this branching continues with further conditions.
2. It is used as an application of supervised machine learning.
3. The algorithm itself evaluates particular decision's importance which is called 'feature classifier'.

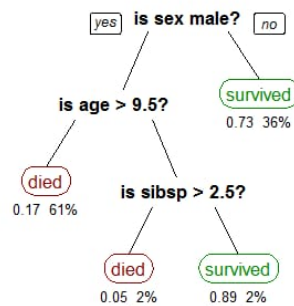


Figure 13: Decision Tree

4. On applying the model with same parameters we get an accuracy of **92.59** % with **250 correct** predictions.

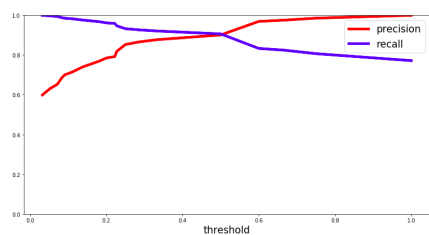


Figure 14: Recall, Precision vs Threshold for Decision Tree

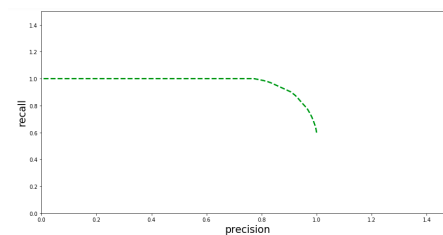


Figure 15: Recall vs Precision for Decision Tree

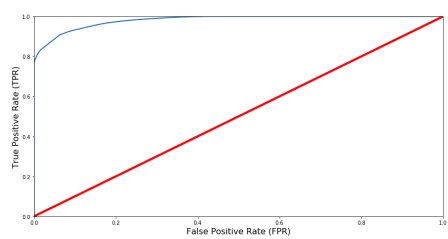


Figure 16: True Positive vs False Positive for Decision Tree

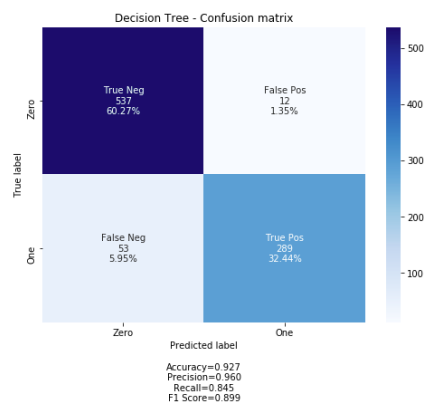


Figure 17: Confusion Matrix for Decision Tree

3.4.4 Random Forest

1. Random Forest is a supervised learning algorithm. It builds a forest of decision trees which are trained using the bagging method.
2. It makes multiple decision trees and then merges them to get more stable and accurate predictions.
3. It brings the extra randomness to the model, by selecting the best feature out of a random set of features rather than selecting from all the features, while splitting the node.

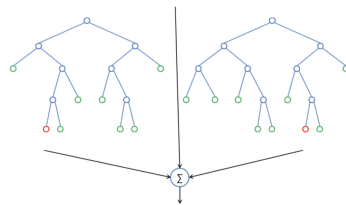


Figure 18: Random Forest

4. It yielded an **accuracy of 92.59%**.

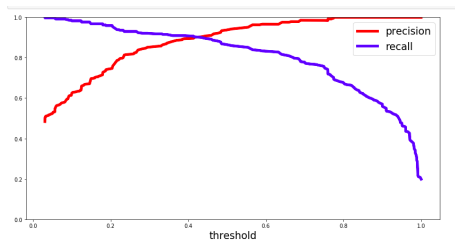


Figure 19: Recall, Precision vs Threshold for Random Forest

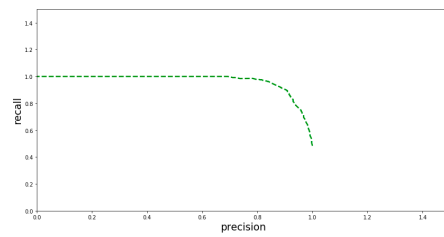


Figure 20: Recall vs Precision for Random Forest

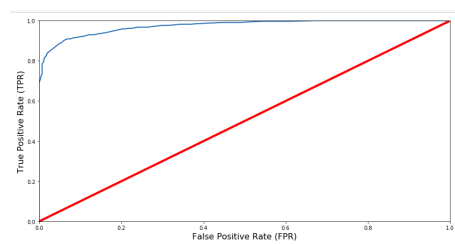


Figure 21: True Positive vs False Positive for Random Forest

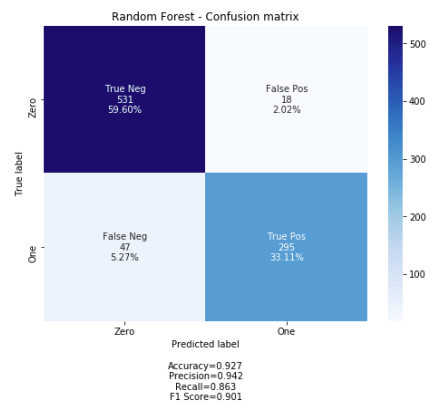


Figure 22: Confusion Matrix for Random Forest

3.4.5 Decision Tree Hypertuning

1. The performance of decision tree can be increased by using pruning. It involves removing branches that make features of low importance.
2. In order to implement best possible model of decision tree ,we can focus on two main hyper parameters:
 - (a) Max Depth- This is the max no of child nodes that can grow out until the tree is cut off.
 - (b) Min Sample Leaf- This is min no. of samples or data points that is required to represent the leaf node.
3. The decision tree has been hyper tuned where random state = 1, max depth = 7 and min-sample-split = 2. i.e. Tree up to 7 levels and 2 children max.
4. It is easier to visualise a decision tree with only 7 levels.
5. This algorithm gave us an **86.86% accuracy**.
6. The accuracy is less than the normal decision tree because we have 13 attributes, and hence 13 level are needed to split on each attribute, while we tuned our Tree to 7 levels only.

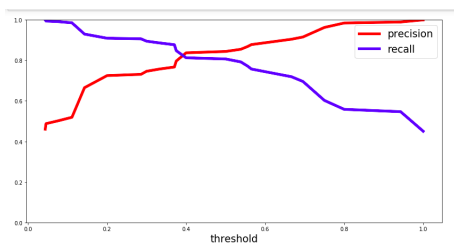


Figure 23: Recall, Precision vs Threshold for Tuned Decision Tree

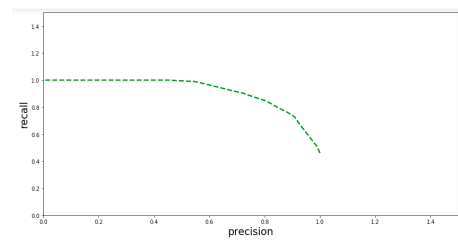


Figure 24: Recall vs Precision for Tuned Decision Tree

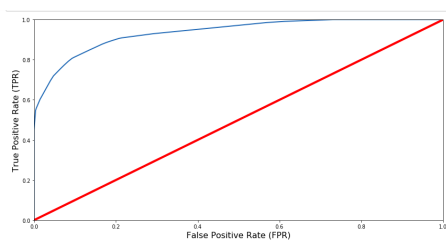


Figure 25: True Positive vs False Positive for Tuned Decision Tree

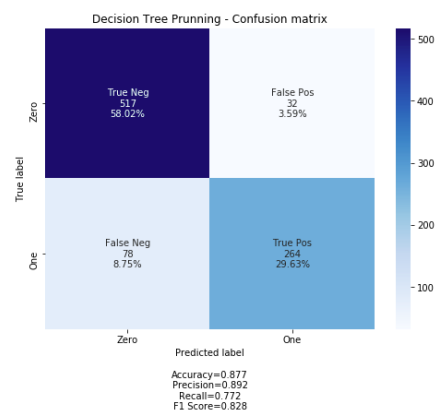


Figure 26: Confusion Matrix for Tuned Decision Tree

3.4.6 Support Vector Machines

1. Support vector machines algorithm is like a knife used to separate data.
2. Each data item is plotted as a point in n-dimensional space with value of each feature being the value of a particular coordinate.
3. Classification is performed by finding a hyperplane that discerns the two classes very well.
4. It has a unique feature to ignore the outliers, hence it is robust to it.

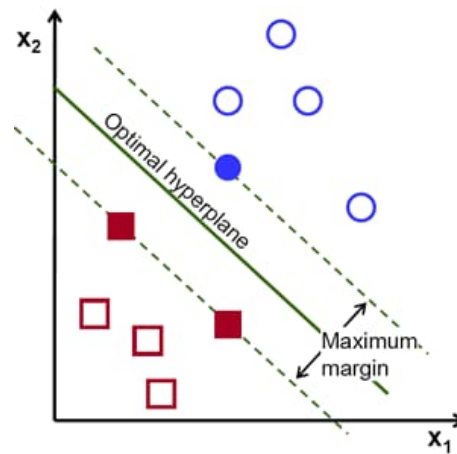


Figure 27: Plane for classification

5. The algorithm yielded us an **accuracy of 81.36%**.

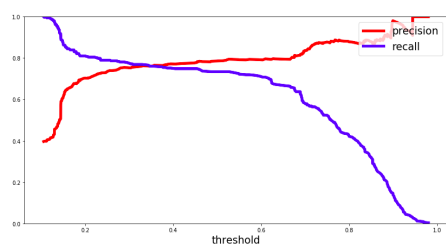


Figure 28: Recall, Precision vs Threshold for Support Vector Machine

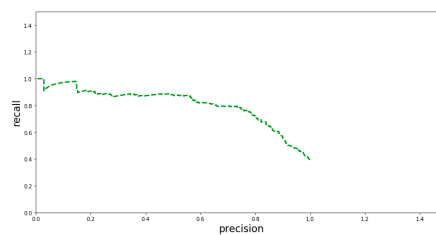


Figure 29: Recall vs Precision for Support Vector Machine

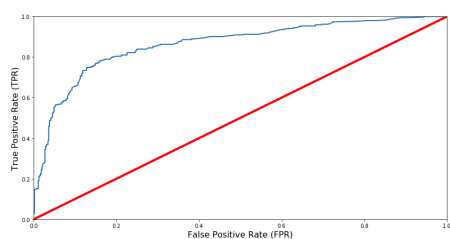


Figure 30: True Positive vs False Positive for Support Vector Machine

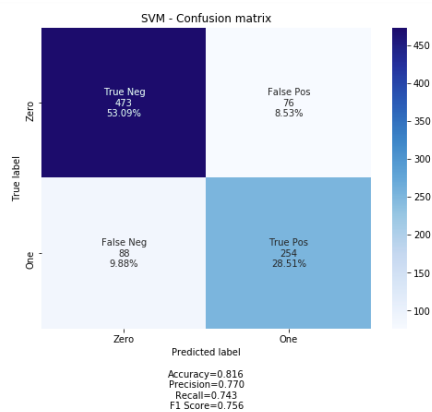


Figure 31: Confusion Matrix for Support Vector Machine

3.4.7 Stochastic Gradient Descent

1. The word ‘Stochastic’ means a system or a process that is linked with random probability.
2. In a typical Batch Gradient Descent, a batch of samples is taken at each iteration to find global minima. However this becomes computationally expensive when we have huge data.
3. This problem is solved by stochastic gradient descent where at each iteration a batch of size 1 is used for applying gradient decent.

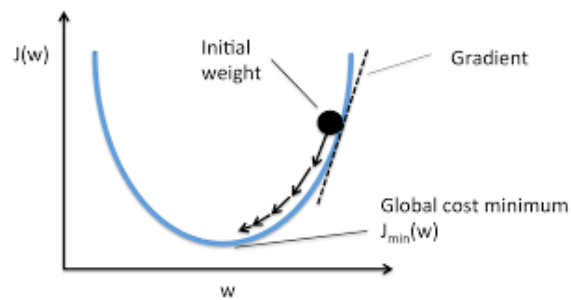


Figure 32: Stochastic Gradient Descent

4. The algorithm yielded us an **accuracy of 66.32%**.

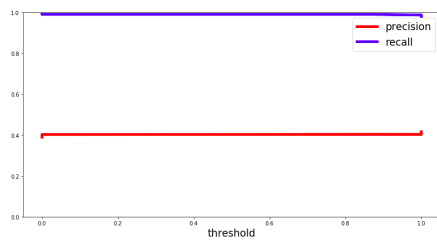


Figure 33: Recall, Precision vs Threshold for Stochastic Gradient Descent

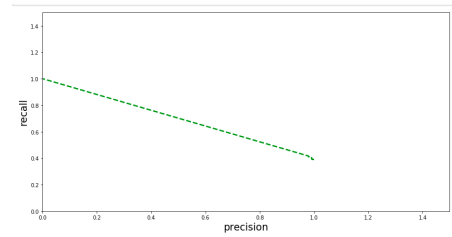


Figure 34: Recall vs Precision for Stochastic Gradient Descent

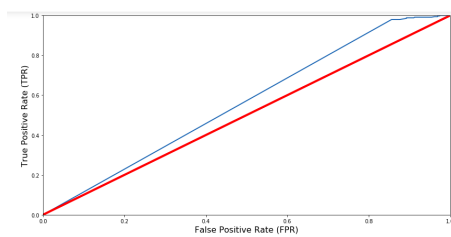


Figure 35: True Positive vs False Positive for Stochastic Gradient Descent

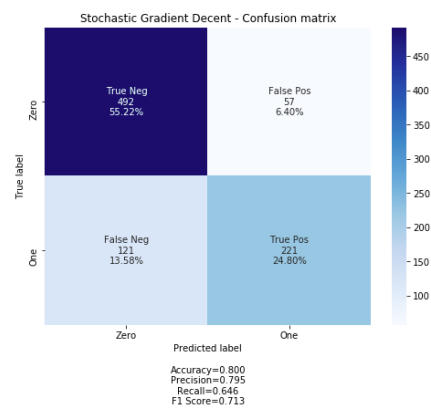


Figure 36: Confusion Matrix for Stochastic Gradient Descent

3.4.8 Perceptron

1. It is the simplest form of a neural network which consists of single neuron for computation.
2. There are no hidden layer. The Input and Output layers are same.
3. Perceptron takes an input and finds the weighted sum then returns 1 only if the sum exceeds the threshold value.

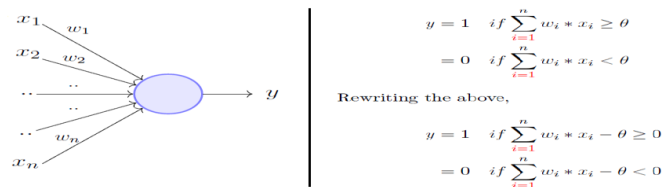


Figure 37: Perceptron

4. The algorithm yielded us an **accuracy of 81.59%**.

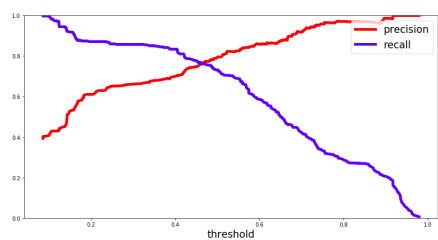


Figure 38: Recall, Precision vs Threshold for Perceptron

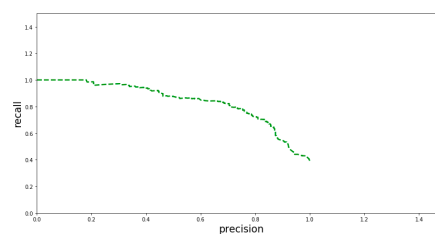


Figure 39: Recall vs Precision for Perceptron

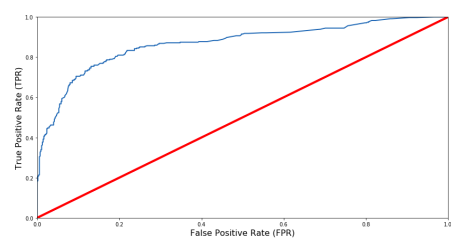


Figure 40: True Positive vs False Positive for Perceptron

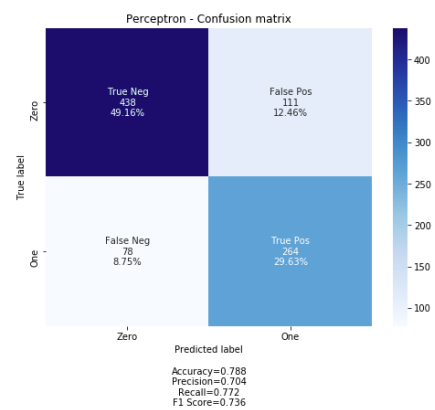
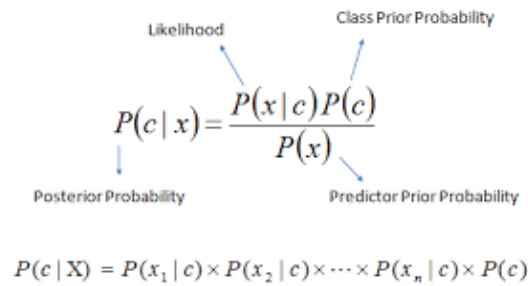


Figure 41: Confusion Matrix for Perceptron

3.4.9 Naive Bayes

1. A Naive Bayes classifier is a probabilistic machine learning model that's used for classification related problems.
2. This algorithm is based on Bayes Theorem.


$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Posterior Probability

Predictor Prior Probability

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

Figure 42: Naive Bayes classifier

3. The main disadvantage of this algorithm is that it considers all features as independent so it cannot learn the dependencies between the features.
4. The algorithm yielded us an **accuracy of 77.2%**.

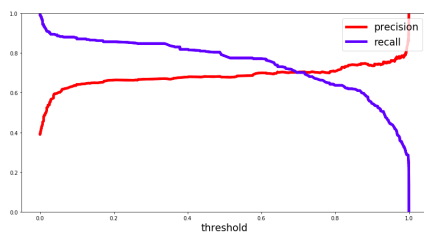


Figure 43: Recall, Precision vs Threshold for Naive Bayes

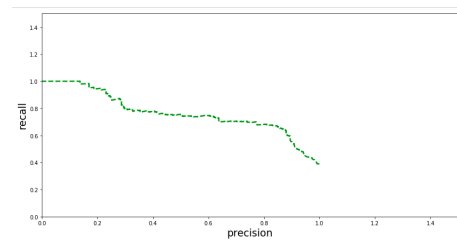


Figure 44: Recall vs Precision for Naive Bayes

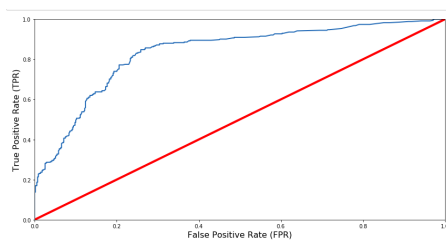


Figure 45: True Positive vs False Positive for Naive Bayes

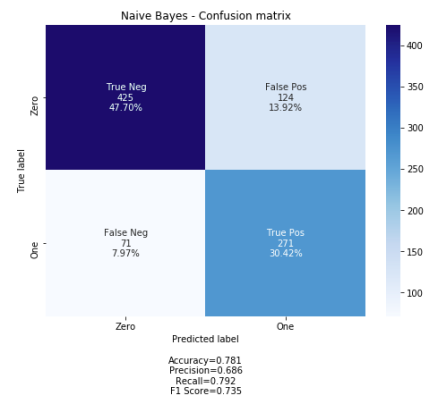


Figure 46: Confusion Matrix for Naive Bayes

3.4.10 Stacking

1. This is **the proposed algorithm**. We have created it using the ensemble learning technique.
2. Stacking is one of the ensembler models where primary objective is to increase mean accuracy as well as to reduce high variance.
3. Stacking incorporates characteristics of both bagging and boosting.
4. Stacking consists of two or more classifiers which act as level-one classifiers. The results from these classifiers act as input for a meta-classifier. This meta classifier gives final prediction.

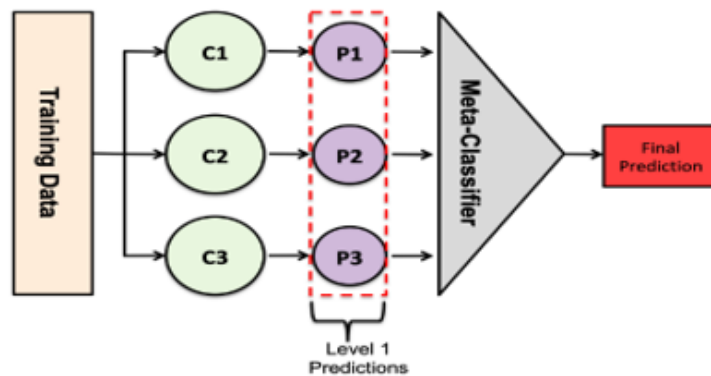


Figure 47: The Stacking Classifier

5. Our Stacking model has 5 level-one classifiers- Random forest, Decision Tree, Extreme Gradient Boosting, Gradient Boosting Machine and Bagging. Meta-classifier used is Logistic Regression.
6. It gives us the **accuracy of 94.17%**.

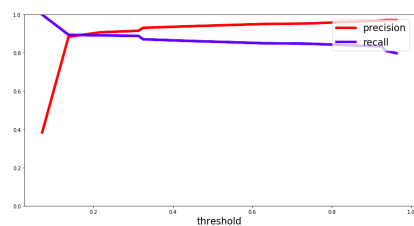


Figure 48: Recall, Precision vs Threshold for Stacking

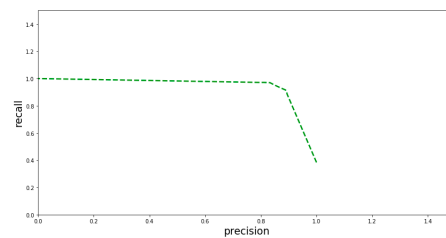


Figure 49: Recall vs Precision for Stacking

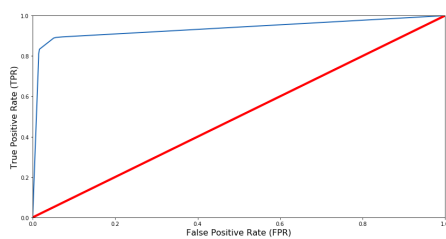


Figure 50: True Positive vs False Positive for Stacking

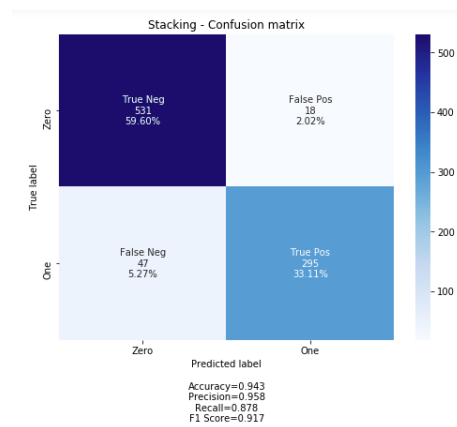


Figure 51: Confusion Matrix for Stacking

Chapter 4

Results And Analysis

4.1 Results

Score	Model
94.167593	Stacking
92.592593	Random Forest
92.592593	Decision Tree
86.868687	Decision Tree with Prunning
81.593715	Perceptron
81.369248	Support Vector Machines
81.369248	Logistic Regression
81.032548	KNN
77.216611	Naive Bayes
66.329966	Stochastic Gradient Decent

Figure 52: The accuracy of various models

1. We see that from all of the existing algorithms , Decision Tree and random Forest give the best performance i.e. 92.59 percent.
2. Naive bayes gives poor performance, reason being the assumption of Naive Bayes that all features contribute equally and are independent to each other which is not in our case as seen in correlation table.
3. Stochastic Gradient Decent inspite of one of the good algorithms, fails here due to under fitting and its more linear nature.
4. The hyper-parameter tuned Decision Tree gives accuracy less than the untuned Decision Tree. Tuned model is over-fitted over training data. Moreover this model is more complex as well takes more time in training as well as in making predictions.
5. Our Stacked model performs extremely well over test data with a accuracy over 94.1675 percent. This is the best mean accuracy achievable so far.

<matplotlib.axes._subplots.AxesSubplot at 0x1a157c1e10>

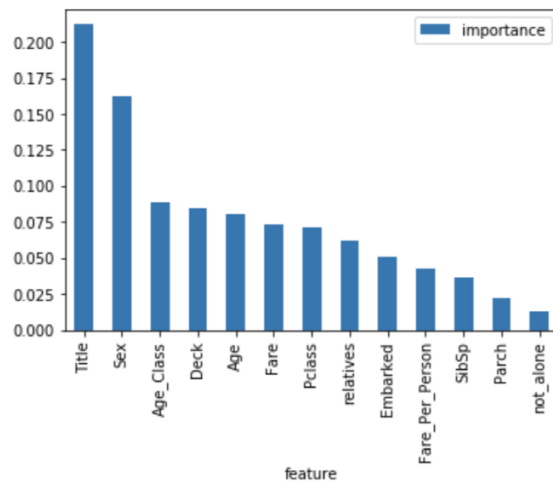


Figure 53: The Importance of various features

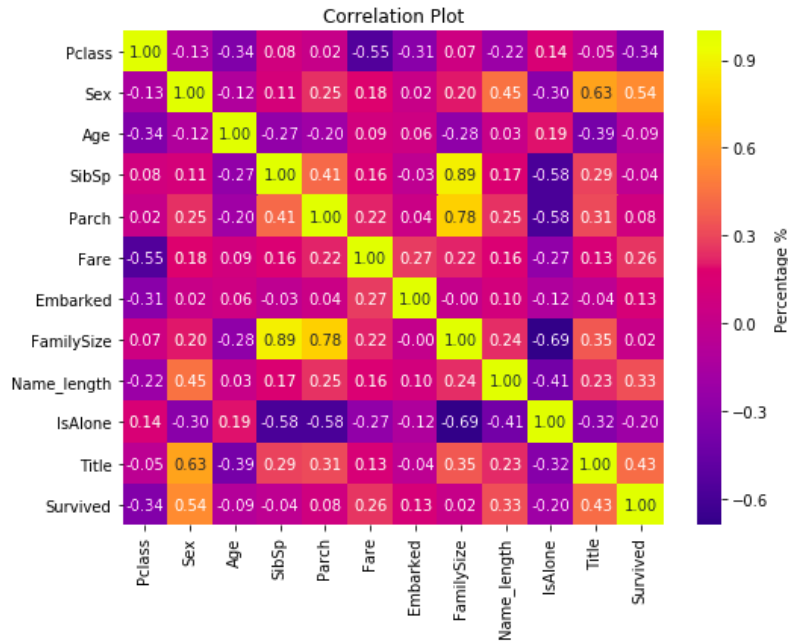


Figure 54: Correlation plot

4.2 Analysis

1. The analysis showed that the **attributes Age, Class of compartment and Gender** are the important features to the model building.
2. Further analysis shows that the majority of passengers who survived occupied the first class.

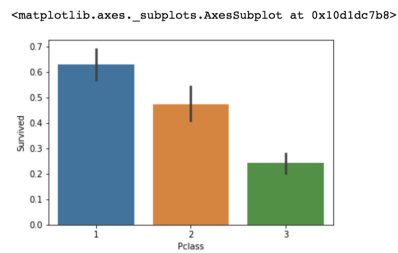


Figure 55: Survivors vs Class of Compartment

3. The majority passengers surviving were of age group between 20-30. Further,

the majority of the men who survived were between 18-30 years of age. While the women who survived ranged between 14-40 majorly.

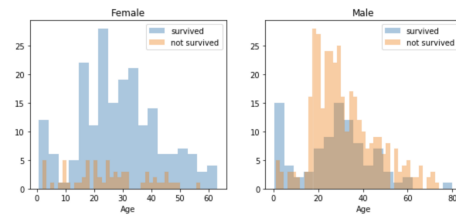


Figure 56: Survivors vs Age of Compartment

4. The passengers surviving were majorly female. The gender is found to me the most important feature for classification.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

1. The comprehensive research gives us a result with decision tree having the highest score with 92.82% correct predictions and lowest false discovery rate.
2. The research also aware us of the features that are highly relevant to the prediction of survival of a passenger, with gender being the feature with the highest importance.
3. The correlation between factors first evaluated using a basic formula was also justified in some cases while being defied in others.
4. An algorithm has been proposed which has a better accuracy than all the existing models.
5. The accuracy of various complex models was relatively low for the data. This is because we had only 891 instances of the data, which lead to over-fitting in most of the cases.

5.2 Applications

1. We can use the model both for predictive and classification purposes.
2. We can predict the missing values of attributes and even the events that may happen in future.
3. The predictive model can be used to predict chances of survival and can be used to prevent any mishap in future.
4. The decision tree gave the best accuracy, hence we know which model works the best for such data.

5.3 Future Work

1. Future work includes using other algorithms like K means, gradient boosting, adaboost, further hyper tuning the decision tree algorithm and even using advanced neural networks as well as Reinforcement learning.
2. Validating other techniques like assigning feature importance, introducing a new feature, a more robust pre-processing could improve the accuracies and may yield different results for different algorithms.

References

- [1] DISASTER, C. T.-M. L. F. Eric lam stanford university.
- [2] EKINCI, E., OMURCA, S. İ., AND ACUN, N. A comparative study on machine learning techniques using titanic dataset. In *7th International Conference on Advanced Technologies* (2018), pp. 411–416.
- [3] FARAG, N., AND HASSAN, G. Predicting the survivors of the titanic kaggle, machine learning from disaster. In *Proceedings of the 7th International Conference on Software and Information Engineering* (2018), pp. 32–37.
- [4] SINGH, A., SARASWAT, S., AND FAUJDAR, N. Analyzing titanic disaster using machine learning algorithms. In *2017 International Conference on Computing, Communication and Automation (ICCCA)* (2017), IEEE, pp. 406–411.