BIG DATA ANALYTICS FINAL PRESENTATION

TITANIC PROJECT

Group 5
Shubham Mehta
Sinem Özeroğlu
Astrid Junco Sommer
Alvaro Taix
Vianney Lacroix

Objective

- Check the “test set” against the “trained model” prepared with training set data and;

- Calculate the Accuracy, Precision and Recall for the trained model.

CONTENT

01.

Data Preparation

What did we clean or eliminate in data and why?

02.

Models used

Which training models did we use?

03.

Our results

What were our results in each model?

04.

Take-aways

What did we learn from “Titanic Project”?

Key observations from the dataset

- For the purposes of analysis, we combined the training and test data set with total values of 1309.
- We obtained that 38% people in the total data set - survived.
- This dataset was 58% of the total number of passengers onboard Titanic.
- Sibsp=0 for 891 passengers and Parch=0 for 1002 passengers. This was posing challenge of unbalanced data and making it difficult to find correlation for these parameters.
- Names were unique.
- Males formed 64.4% of the total cohort.

Key observations from the dataset

- Possibility of sub-categorization in males among those with Mr. and Master. titles due to significant median age variation.

- Females had a very high survival rate of approx 74%.

- Fares varied significantly and most of the cabin values were missing and had several duplicates.

- S-Southampton port had the most number of embarkments.

- P-class of 3 consisted of majority of passengers and most of them couldn’t survive. While majority of people in P-class=1 survived.

1- DATA CLEANING

Cleaning

1- Checking for empty/missing values

Values missing in parameters namely fare, age, cabin, ticket number.

2- Trying to fill empty values

Filling the values with mean and median values for age & fare.

3- Eliminating specific parameters

Some columns and parameters were taken out as they could not be filled reasonably like embarkment and cabin number.
Other columns that showed no correlation with survival prediction like Passenger ID, name were also taken out.

4- Creation of new parameters

New parameters with entities like Mr.,Mrs. etc were created.

5- Checking for duplicates

None duplicate was found in name while only a few were found in values like ticket number, cabin number and fares.

6-Checking for outliers

No outliers were found.

2- TRAIN MODELS

Three different sets of model

1- Mean values in age

6 parameters were used; pclass, sex, age, sibsp,parch and embarkment

2- Median values in age

One more parameter added: prices

3-Subcategorization of males

Mr. and Master

3- RESULTS

01. First Set

DNN showed highest accuracy at 77.9%, Random forest at 75.5% and decision tree at 75.1%.

02. Second Set

With the parameter of prices included, DNN decreased by 0.15%, decision tree also reduced by 1.9% but random forest showed increase in accuracy by 0.3%.

03. Third Set

DNN retained the same accuracy from the previous iteration at 77.75%, but random forest and decision tree both showed reduction in accuracy to the lowest levels amongst the three iterations done so far. So we obtained that DNN with initial six parameters provides the best accuracy.

MEDIAN	SUBCATEGORIZATION
● Median values to fill the missing cells for ages; ● Picked the most accurate model of DNN to see if the accuracy could be improved; ● So the model contained six initial parameters along with titles but not fares (which had led to reduction of accuracy of DNN model); ● We found out that accuracy further improved for DNN model by 0.5% to reach 78.4% ● The precision and recall for this model is 72.3% and 69.6% ● Checked with Decision tree and Random forest	● Sub-categorised males into Mr. and Master ● Clear distinction in the median and mean values of these two sets ● The accuracy, precision and recall for this model turned out to be 76.5%, 69.2% and 68.3% respectively ● Checked with decision tree and random forest afterwards

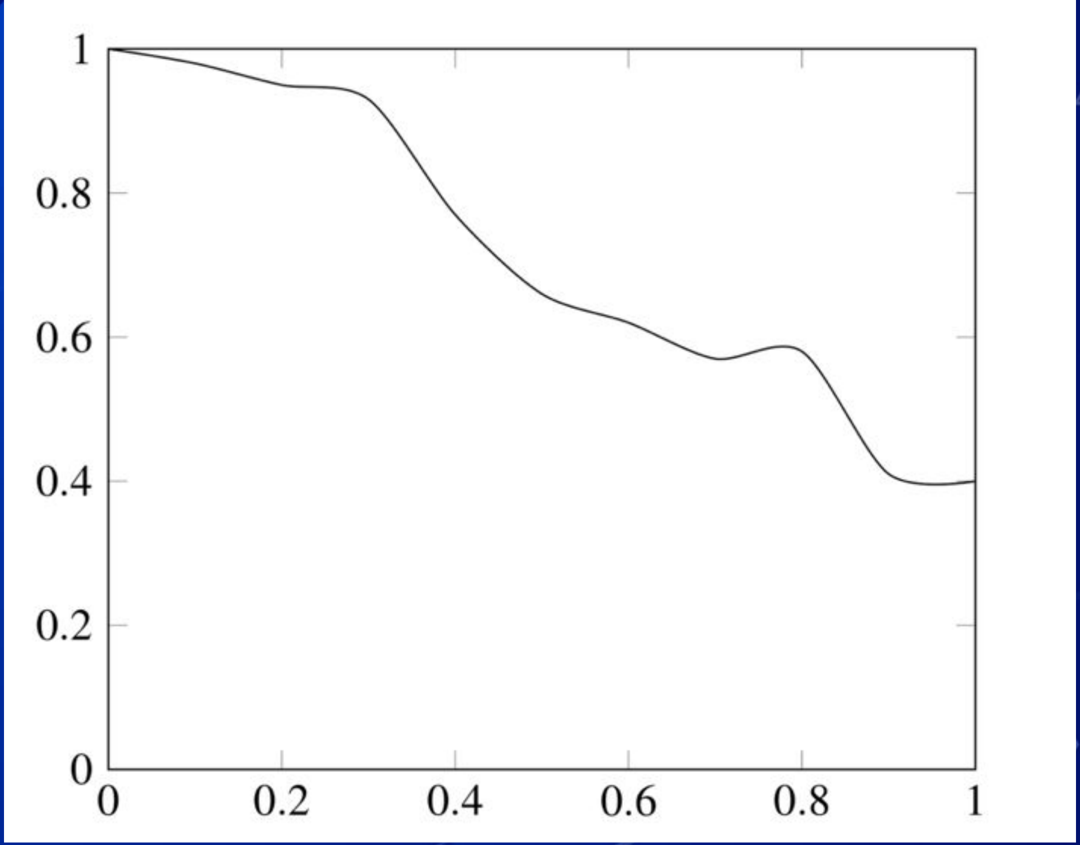
	Decision Tree				Random Forest				Deep Neural Network			
	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
1. Average with Six parameters	0.751	0.673	0.664	0.668	0.755	0.679	0.670	0.674	0.779	0.75	0.626	0.682
2. Six parameters+Fares	0.732	0.649	0.632	0.64	0.758	0.699	0.632	0.663	0.777	0.748	0.62	0.678
3. Six parameters+Fares+ Titles	0.717	0.626	0.626	0.626	0.746	0.675	0.632	0.652	0.777	0.712	0.689	0.7
4. Median +Titles(no Fares)	0.722	0.619	0.689	0.652	0.727	0.642	0.626	0.633	0.784	0.723	0.696	0.709
5. Median +Titles(no Fares) with male sub-categorization	0.732	0.63	0.702	0.664	0.751	0.673	0.664	0.668	0.765	0.692	0.683	0.687

Confusion Matrix for the most accurate model

Predicted ➡ Actual ⬇	Survived	Not Survived	
Survived	110	48	Recall=110/158=0.696
Not Survived	42	218	
	Precision=110/152=0.723		Accuracy=328/418=0.784

Relation between Precision and Recall

Precision



Recall

Source: Towards data science

4- TAKEAWAYS

Deep Neural Network

- DNN turned out to be the best model with highest accuracy at 78.4%
- Since our training and test set had mix of categorical and numerical values-Complex problem
- Back-propagation in Deep neural network works to ensure that appropriate weights are assigned to each parameter
- Works perfectly with high-dimensional data and high cardinality outcomes where there are a number of discrete categories available in the dataset.

Random forest

- The accuracy of the model increased with higher sub-categorization within the dataset.
- The longer the tree, the better is the prediction accuracy.

Decision Tree

- The accuracy kept on decreasing as we added more parameters to our model in training and test set.
- This is because decision tree doesn't work well if lots of parameters are involved.
- Not competitive amongst supervised learning approaches in terms of prediction accuracy.

Precision and Recall

- Not necessarily, the value of precision increases as recall decreases and vice-versa.
- F1 ensure that extreme values of precision or recall are balanced with harmonic mean.

Would you have survived?

Names	Pclass	SibSp	Parch	Embarkment	Age	Sex	Survived
Shubham	3	1	0	S	24	M	0
Astrid	2	1	0	S	22	F	1
Sinem	2	0	1	S	23	F	1
Vianney	3	1	1	S	23	M	0
Alvaro	1	1	0	S	22	M	1

Thanks !