Spaceship Titanic: Binary Classification

Final Project on Fundamentals of Data Science (FDS)

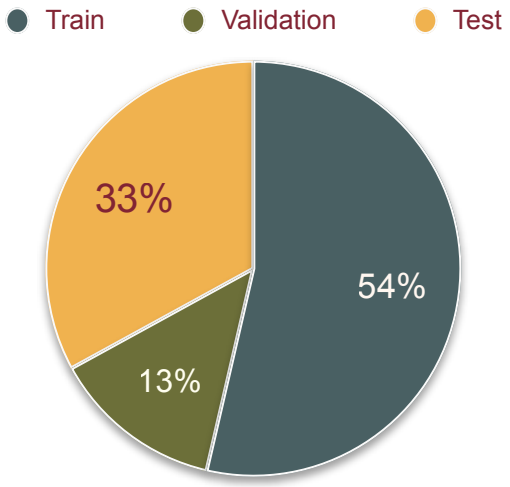
MS Data Science

AY: 2023-2024

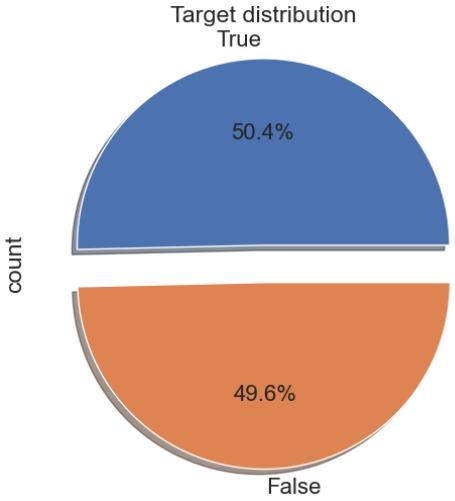
Ambar Chatterjee, Himel Ghosh, Paul Jèzèquel, Mayis Atayev

SAPIENZA
UNIVERSITÀ DI ROMA

Data: Source and Distribution



- Source: Kaggle [https://www.kaggle.com/competitions/spaceship-titanic/data]
- train.csv (8693x14), test.csv (4277x13)



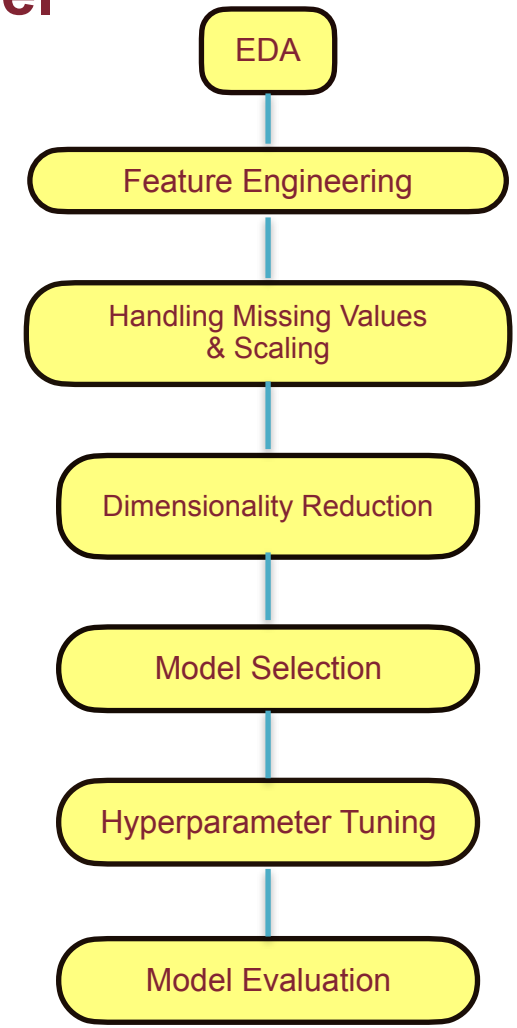
Original Data:

	PassengerId	HomePlanet	CryoSleep	Cabin	Destination	Age	VIP	RoomService	FoodCourt	ShoppingMall	Spa	VRDeck	Name	Transported
0	0001_01	Europa	False	B/0/P	TRAPPIST-1e	39.0	False	0.0	0.0	0.0	0.0	0.0	Maham Ofracculy	False
1	0002_01	Earth	False	F/0/S	TRAPPIST-1e	24.0	False	109.0	9.0	25.0	549.0	44.0	Juanna Vines	True
2	0003_01	Europa	False	A/0/S	TRAPPIST-1e	58.0	True	43.0	3576.0	0.0	6715.0	49.0	Altark Susent	False
3	0003_02	Europa	False	A/0/S	TRAPPIST-1e	33.0	False	0.0	1283.0	371.0	3329.0	193.0	Solam Susent	False
4	0004_01	Earth	False	F/1/S	TRAPPIST-1e	16.0	False	303.0	70.0	151.0	565.0	2.0	Willy Santantines	True

Data Pipeline & Baseline Model

Models in Consideration:

Random Forest
Gradient Boosting
Neural Network
Stacked

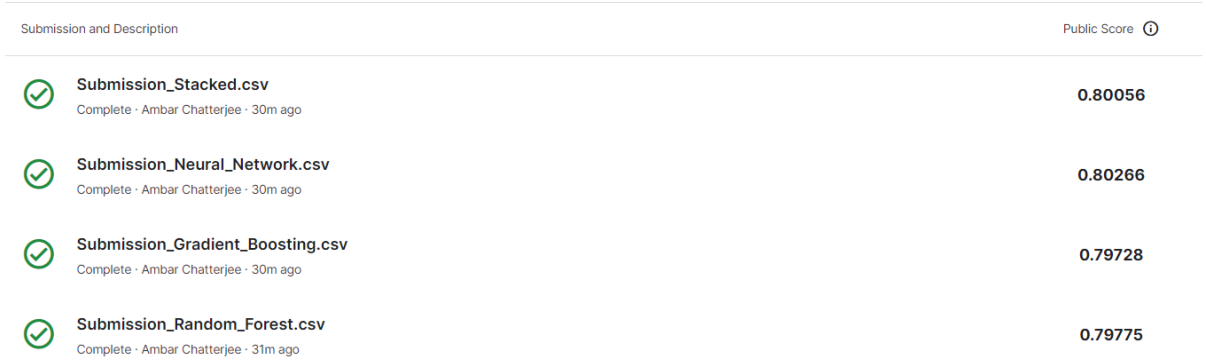


Discussion of Results:

Model Evaluation

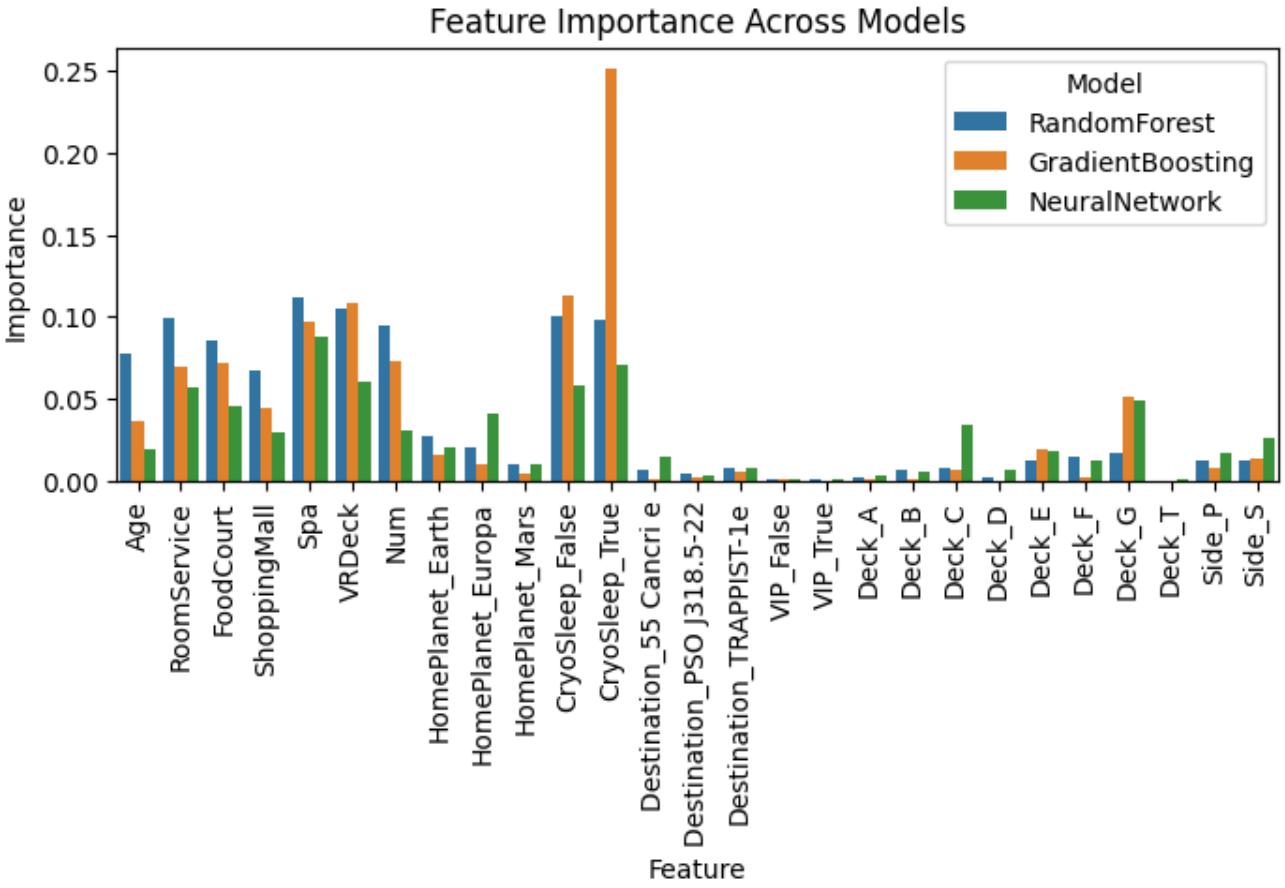
	Accuracy	Precision	Recall	F1 Score	ROC AUC
Random Forest	0.801807	0.783257	0.829891	0.805900	0.891429
Gradient Boosting	0.801807	0.783257	0.829891	0.805900	0.895702
Neural Network	0.795783	0.822667	0.749696	0.784488	0.892720
Stacked	0.803614	0.784000	0.833536	0.808009	0.890130

Our Submission Score on Kaggle



Discussion of Results:

Feature Importance

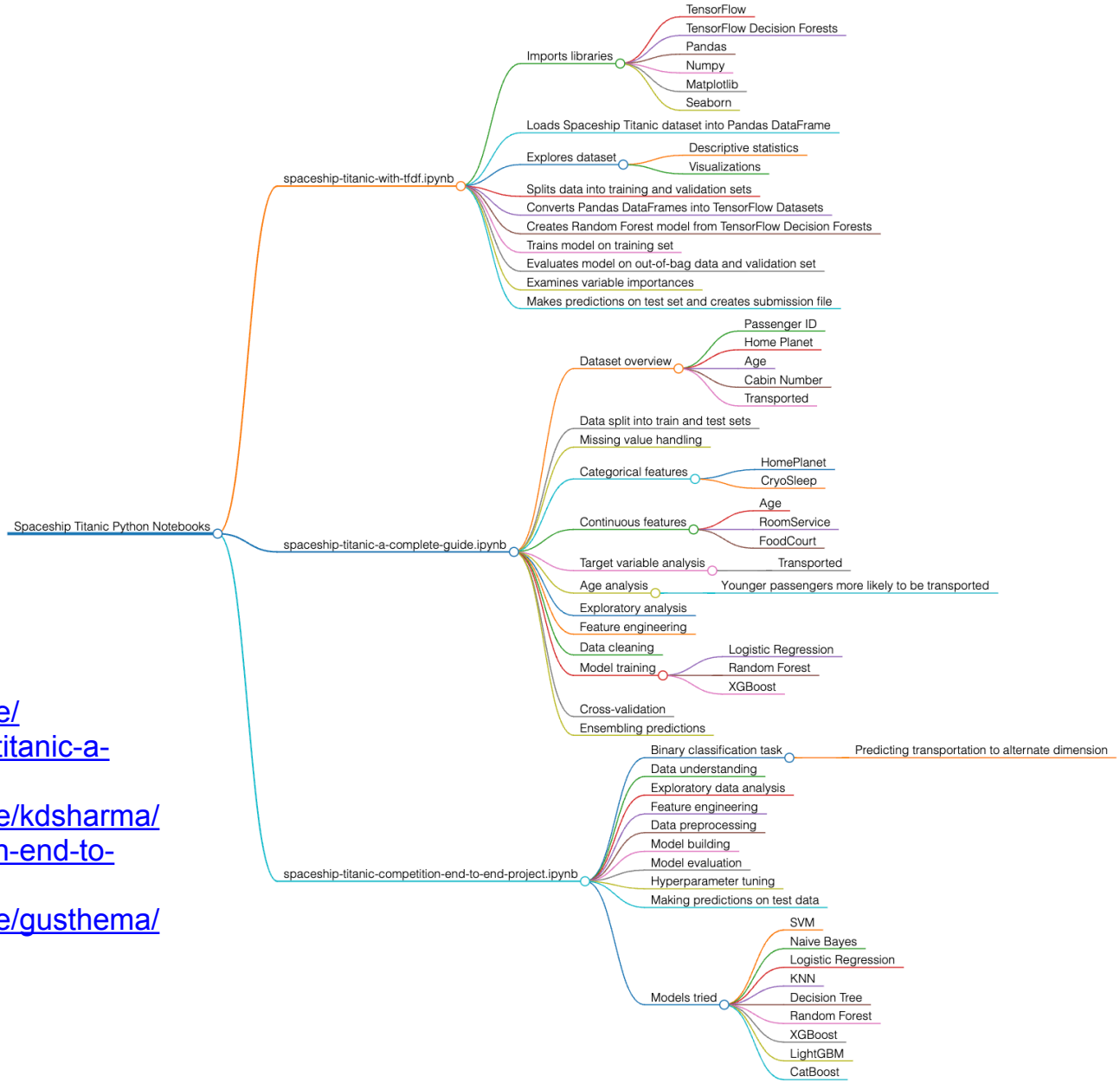


Findings:

We compared all these models based on various indicators and found out that:

- Neural Networks showed best Precision score
- Gradient boosting was the best model based on the ROC AUC score
- Stacked model performed better according to Accuracy, Recall & F1 score
- Models don’t share the same important features
- Handling missing values decreased score
- Dimension reduction lead to slight decrease in score
- Oversampling of the training data resulted in decreased score
- Minimal Feature engineering got us higher scores compared to advanced feature engineering.

Literature Review:



References:

1. https://www.kaggle.com/code/samuelcortinhas/spaceship-titanic-a-complete-guide
2. https://www.kaggle.com/code/kdsharma/spaceship-titanic-competition-end-to-end-project
3. https://www.kaggle.com/code/gusthema/spaceship-titanic-with-tfdf

Thank You

Q&A!

SAPIENZA
UNIVERSITÀ DI ROMA