

Feature Selection for Machine Learning

Emmanuel AFRIFA

Supervisor: Tony TONA LANDU

*African Institute for Mathematical Sciences
(AIMS-GHANA)*



Table of Contents

1. Introduction

2. Methodology

3. Results and Discussions

4. Conclusion

Introduction

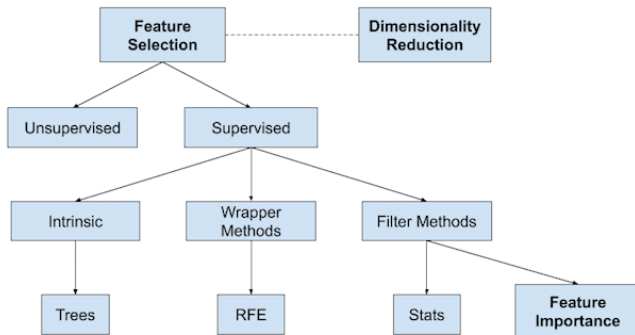


Figure 1: Overview of feature selection techniques

Methodology

- The dataset used was the titanic dataset (from Kaggle) and it has 891 observations and 11 features and 1 target which is the *survival* column.

Methodology

- The dataset used was the titanic dataset (from Kaggle) and it has 891 observations and 11 features and 1 target which is the *survival* column.
- Language and Libraries:
 - Python programming language
 - Pandas
 - Matplotlib and Seaborn
 - Scikit-Learn
 - Xgboost

Methodology

- The dataset used was the titanic dataset (from Kaggle) and it has 891 observations and 11 features and 1 target which is the *survival* column.
- Language and Libraries:
 - Python programming language
 - Pandas
 - Matplotlib and Seaborn
 - Scikit-Learn
 - Xgboost
- Feature selection Methods:
 - Random Forest
 - Recursive Feature Elimination
 - Xgboost

Methodology

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S

Figure 2: Overview of dataset

Methodology

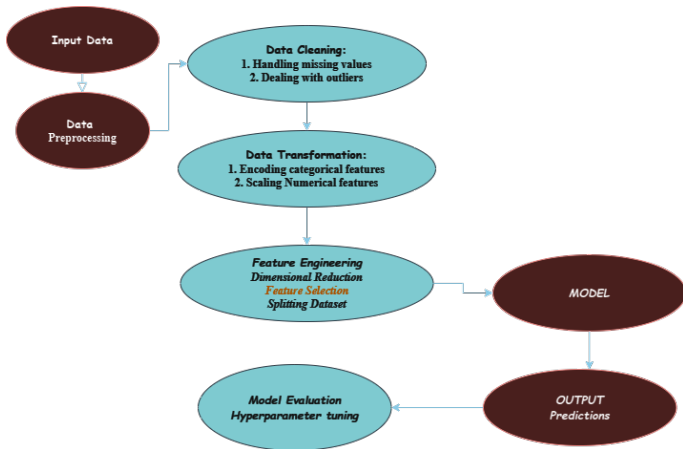


Figure 3: Machine Learning Workflow



Results and Discussions

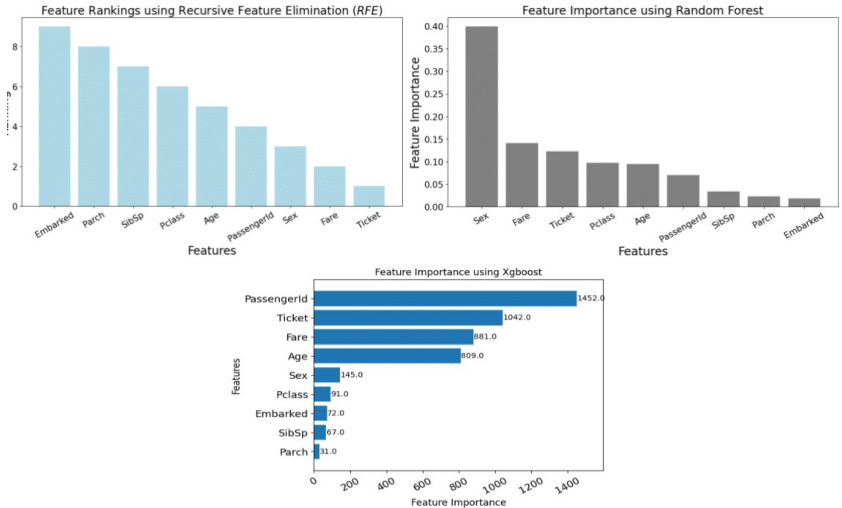


Figure 4: Comparing the three feature selection methods used

Results and Discussions

Feature Selection Method	Features	Accuracy
Without Feature Selection	PassengerId, Pclass, Sex, Age, SibSp, Parch, Ticket, Fare, Embarked	74.30%
RFE Feature Selection	Embarked, Parch, SibSp, Pclass, Age, PassengerId, Sex	79.89%
Random Forest	Sex, Fare, Ticket, Pclass, Age, PassengerId, SibSp	80.44%
Xgboost	PassengerId, Fare, Age, SibSp, Pclass, 'Sex', Ticket	80.44%

Table 1: Different feature selection methods and accuracy.

Conclusion

- In this study, we reviewed the different feature selection methods and investigated the impact of three methods (random forests, recursive feature elimination, and Xgboost) on the accuracy of the model.

Conclusion

- In this study, we reviewed the different feature selection methods and investigated the impact of three methods (random forests, recursive feature elimination, and Xgboost) on the accuracy of the model.
- Using Feature selection increased the accuracy of the model in making predictions.






Conclusion

- In this study, we reviewed the different feature selection methods and investigated the impact of three methods (random forests, recursive feature elimination, and Xgboost) on the accuracy of the model.
- Using Feature selection increased the accuracy of the model in making predictions.
- Although the different methods gave different feature rankings, they all gave a higher accuracy than the model built without any feature selection.

Conclusion

- In this study, we reviewed the different feature selection methods and investigated the impact of three methods (random forests, recursive feature elimination, and Xgboost) on the accuracy of the model.
- Using Feature selection increased the accuracy of the model in making predictions.
- Although the different methods gave different feature rankings, they all gave a higher accuracy than the model built without any feature selection.
- The random forest and Xgboost methods give the highest accuracy for this dataset.

References

-  Jason Brownlee (2020). How to Choose a Feature Selection Method For Machine Learning
-  Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26, p. 13). New York: Springer.
-  Ndung'u, R. N. (2022). Data Preparation for Machine Learning Modelling.
-  Michal Oleszak (2023). Feature Selection Methods and How to Choose Them
-  Titanic - Machine Learning from Disaster

