**UNIVERSITY OF SAINT THOMAS**

**Software Engineering and Data Science**

**Professor: Dr. Chih Lai**

**SEIS 763: Machine Learning, Summer 2024**

Final Project

Titanic: Machine Learning for Disasters

**Students:**

Darwin Stahlback

Montserrat Aguilar Rios

Owen Ricker

Zachary Ragone

1. Introduction

The idea for this project came directly from a Kaggle competition. “The competition is simple: use machine learning to create a model that predicts which passengers survived the Titanic shipwreck.” The sinking of the Titanic is one of the most infamous and tragic events in history. 1,502 of the 2,224 passengers and crew perished in the Titanic’s maiden voyage after hitting an Iceberg in the North Atlantic Ocean. This has led to numerous changes in ship regulations, deep dive explorations, and even a famous James Cameron movie adaptation. Today, Kaggle is running a challenge to build a predictive model to answer the question: “what sorts of people were more likely to survive?’ using passenger data (i.e. name, gender, socio-economic class, etc.).” We answer that question using multiple Machine Learning modeling techniques: Lasso logistic regression, Radial Base Function Kernel Support Vector Machine and three tyes of Decision Trees (Basic Decision Tree, Pruned Decision Tree and Random Forest).

1. Description of Data Source

Data is direct from a [**Kaggle competition**](https://www.kaggle.com/competitions/titanic). The competition details the features of passengers onboard and whether they survived the fateful crash. We are given a Training data csv to train our model on and a separate test.csv to then compare against and score our model. The test data set does not have the survival field to avoid cheating during the Kaggle competition.  To mitigate this, we have split the training set provided by Kaggle, in order to evaluate on top of it.

1. Records and Attributes

**Source:** [**DATA**](https://www.kaggle.com/competitions/titanic/data)

* # of records: 891 (train) + 418 (test) = 1,309 records in total
* # of attributes: 10

|  |  |  |
| --- | --- | --- |
| **Variable** | **Definition** | **Key** |
| survival | Survival | 0 = No, 1 = Yes |
| pclass | Ticket class | 1 = 1st, 2 = 2nd, 3 = 3rd |
| sex | Sex |  |
| Age | Age in years |  |
| sibsp | # of siblings / spouses aboard the Titanic |  |
| parch | # of parents / children aboard the Titanic |  |
| ticket | Ticket number |  |
| fare | Passenger fare |  |
| cabin | Cabin number |  |
| embarked | Port of Embarkation | C=Cherbourg, Q=Queenstown, S=Southampton |

**Notes on the attributes:**

**pclass** A proxy for socio-economic status (SES)

1st = Upper, 2nd = Middle, 3rd = Lower

**Age** Age is fractional if less than 1.

If the age is estimated, is it in the form of xx.5

**sibsp** First set of family relations described in this way:

Sibling = brother, sister, stepbrother, stepsister

Spouse = husband, wife (mistresses and fiancés were ignored)

**parch** Second set of family relations described in this way:

Parent = mother, father

Child = daughter, son, stepdaughter, stepson

Some children travelled only with a nanny, therefore parch=0 for them.

1. General Dataset Statistics

**Training Data Set Numeric Statistics                                    Test Data Set Numeric Statistics**

A table of numbers and numbers

Description automatically generated                  A table with numbers and text

Description automatically generated

**Training Data Set Categorical Statistics**

A black text on a white background

Description automatically generated

A white background with black text

Description automatically generated

A black and white text

Description automatically generated

A white background with black text

Description automatically generated

**Test Data Set Categorical Statistics**

A white background with black text

Description automatically generated

A black and white text

Description automatically generated

A screenshot of a computer

Description automatically generated

**Gender Submission Data Set Categorical Statistics**

A black text on a white background

Description automatically generated

1. Tools / methods to use in our study.

Our goal was to use all appropriate classification machine learning (ML) methods/models that we’ve become accustomed to using in class with standardized data. We utilized automated EDA tactics from python libraries like Ydata\_profiling to extract insight and various other plans of attack like SweetViz which generates an \*.html report of every insight and measure of data. These auto-EDA tools help in extracting as many valuable features as possible by showing exactly what and where the useful pieces of data are in this set. We also dove into feature engineering and other sorts of imputation where needed when encountering null values and outliers. When a variable had a significant number of missing values, we imputed with the average or mode of the values. The models we used were: Logistic Regression with a Lasso penalty, SVM (RBF), and Decision Trees (Basic, Pruned, & Random Forest). We calculated Accuracy, Precision, Recall, and F-1 score for each model, as well as ROC AUC. For the decision trees we plotted training vs testing gaps and compared feature importance as well as their confusion matrices. We tried to visualize our results and model evaluations to determine which model performed the best and would be the greatest suit for predicting survival.

1. Problem Description and Study Goal

The dataset's big theme is predicting passenger survival and looking at factors influencing survival outcomes. Our group has come up with topics within these themes.

1. How do different demographics (age, class, gender) influence survival chances? What features (e.g., age, sex, class, fare, number of siblings/spouses aboard, number of parents/children aboard) predict survival?

Questions to pose on how demographic and socio-economic factors impacted survivability:

* Are females more likely to survive compared to males?
* Does age play a role in survival probability?
* Do family relationships (number of siblings/spouses, parents/children) affect survival?
* Is there any correlation between passenger fare and survival?

1. What are the correlations between features, and how do they affect the model?
2. Which machine learning techniques work best for these classification problems? How do different models perform in predicting survival? We would evaluate the performance of our models using metrics accuracy, precision, recall, and F1 score. (Models: logit, SVM, three different DT’s)
3. Modeling, Techniques, and Processes

We used the following **PRE-MODELING TECHNIQUES:**

* Filling/Dropping nulls: There were only 2 null for Embarked so we dropped these entirely.
* Imputed missing values for 'Age' and 'Fare' using the mean.
* The following data conversions were made:
  + Standardization of numeric variables.
  + One-hot encoding for categorical variables.
  + Binarization of 'Parch' and 'SibSP', which was turned into a Yes/No relationship
* Feature Engineering
  + The 'Name' for passengers also included a 'Title' (e.g.: Miss, Mlle, Ms, Lady), so we turned these into a new variable called ‘Title’.
  + 'Cabin' grouped as 'Cabin\_Floor' > 9 floors
  + The following columns were dropped:
    - Ticket, dropped since values are categorical and extremely distinct.
    - Cabin / Gender dropped due to high multi-collinearity with other engineered features.

We used the following **MODELING TECHNIQUES:**

1. **Logistic Regression**

* We applied *Lasso Regularization* with Cs = 30. Why?
  + For the size of our dataset (819 records), 30 values represent a reasonably wide range of regularization strengths.
  + This also allows the model to explore from very strong (small C) to very weak (large C).
  + We provide balance between granularity and computational cost:
  + If there are too few, we would’ve risked missing optimal C. If there are too many, it becomes computationally expensive.
* With this technique, we discovered that the features that have a higher impact in survivability are the following:
  + Title\_Mrs. (3.05) 🡪denotes a woman
  + Title\_Miss (2.57) 🡪 denotes a woman
  + Title\_Master (2.29) 🡪 denotes a child
  + Class\_1 (1.98)
  + Class\_2 (0.99)

Consider that the first two titles ‘Mrs.’, ‘Miss’ denote the gender ‘female’. Additionally, the title Master denotes a child which may be linked to the age of the passenger. Finally, Class\_1 and Class\_2 highlight a possible connection with the ‘fare’ paid by each passenger. Further analysis may prove beneficial, and this will be addressed in the conclusions.

* These are the **scores** obtained using logistic regression:

A screenshot of a computer

Description automatically generatedA graph of a logistic curve

Description automatically generated

1. **Radial Base Function Kernel Support Vector Machine Model (RBF SVM)**

* We used different combinations of C, as small as 0.1 to 200.
* We also ran Gamma for as low as 0.1 to 10.
* The best results for Test and Train Accuracy and F-Score is C = 2.83 and Gamma = 0.9.
* This is the lowest C and Gamma with the best score in all 4 test statistics.

A screenshot of a computer

Description automatically generated

A graph of a function kernel support

Description automatically generatedA screenshot of a computer

Description automatically generated

1. **Decision Tree, Pruned Decision Tree and Random Forest.**
   1. BASIC TREE: We started with a “basic decision tree”. It resulted in an overfitting tree.

A diagram of a network

Description automatically generated

A screenshot of a computer

Description automatically generatedA graph of a line

Description automatically generated

* 1. PRUNED TREE**.** Then we pruned the basic tree, and this is the final PRUNED TREE:

A screenshot of a computer screen

Description automatically generated

A diagram of a graph

Description automatically generated

* 1. RANDOM FOREST**.**

After multiple test analyses, our team implemented the Random Forest model with 200 trees, each up to 10 levels deep. This configuration demonstrated better performance, achieving approximately 80% accuracy in predicting Titanic survivors.

While our analysis showed that the performance gains begin to level off around 50-75 trees, we found that extending to 200 trees offered several key advantages:

1. Good generalization: The larger ensemble of trees enhanced the model's ability to capture subtle patterns in the data, leading to improved performance on unseen data.

2. Reduced variance: With 200 trees, our model benefits from a more diverse set of decision paths, resulting in more stable and reliable predictions across different subsets of the data.

3. Feature importance stability: The increased number of trees provides a more comprehensive view of feature importance, ensuring that our insights into survival factors are more reliable and less prone to random fluctuations.

4. Allows for Future-proofing: As potential additions to our dataset or slight shifts in data distribution are anticipated in the Kaggle competition , the more complex model is better equipped to adapt without significant retraining.

5. Little computational trade-off: Given the relatively small size of the Titanic dataset, the additional computational cost of 200 trees versus 75 was minimal compared to the performance benefits.

A graph of a number of trees

Description automatically generatedA chart of different colors

Description automatically generated

* When comparing all decision trees (basic, pruned and forest), the most important features are the following:
  + Age, Fare, Title Mrs., Title Miss., Class 2, and Class 1.

It is interesting that some of these are also the most important features determined by the Lasso technique described earlier: Title\_Mrs., Title\_Miss, Class\_1, Class\_2. The relationship between these features provides an opportunity for further exploration, as noted in the “Conclusions” section.

A graph of blue and white bars

Description automatically generated

A screenshot of a graph

Description automatically generated

* When looking at the LEARNING CURVES for the decision trees (basic, pruned and forest), we see that:
* **BASIC tree**: There is a large gap between training and CV scores, this indicates overfitting.
* **RANDOM**: The gap between training and CV scores narrows, showing less overfitting than before.
* **PRUNED**: There is a smaller gap between training and CV scores indicates better generalization.

A graph with green and red lines

Description automatically generated

* On the CONFUSION MATRICES for the decision trees (basic, pruned and forest), we see that:
* **Pruned decision tree improved the basic tree:**

↓ Reduced count of false positives, from 19 to 9.

↑ Increased count of true negatives, from 92 to 102.

* **Random Forest model balances the basic and pruned trees:**

↓ Reduced false negatives compared to both (21 vs 24)

↑ Slight increase in true positives (46 vs 43)

False positives (14) fall between the basic (19) and pruned (9) trees.

A graph of a tree confusion matrix

Description automatically generated

* As for the SCORES for all three (basic, pruned and forest), we see that:
* The Pruned Decision Tree offers the best accuracy / precision. So, if precision is most critical, we should select this one.
* However, IF balance with strong recall is needed, the Randon Forest RF is better.

A graph of different colored bars

Description automatically generated

* Finally, looking at the AUC for all three (basic, pruned and forest), we see that:
* Random Forest (green line) IS BEST: AUC of 0.86
* Pruned DT (orange line) is 2ND best: AUC of 0.81
* Basic DT (blue line) WORST: AUC of 0.73

*OVERALL CONCLUSION for DECISION TREES: Random Forest has the highest curve, and a better overall performance. Pruned Decision Tree offered an improvement over Basic Decision Tree, since pruning helped reduce overfitting.*

A graph of a line graph

Description automatically generated with medium confidence

1. Summary of Results

* Only looking at “Accuracy”, the best model is SVM RBF, with the second best being the pruned Decision Tree (see bar graph below).

A graph of a model

Description automatically generated with medium confidence

* Additionally, also considering Precision, Recall, and F-1 score, the best model is also the **RBF SVM**, with the second best being also the **Pruned Decision Tree**, closely followed by the **Random Forest.**

***LOGIT RBF SVM***

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

***DECISION TREES / RANDOM FOREST***

A graph of different colored bars

Description automatically generatedA graph of a line graph

Description automatically generated with medium confidence

1. Conclusions

* Machine Learning provided excellent methods for the Kaggle Titanic dataset, which is imbalanced and has a combination of numerical and categorical variables.
* The most important features that were highlighted by both Lasso and the Pruned Decision Trees were Title\_Mrs., Title\_Miss, Class\_1, Class\_2. This highlights the importance of Gender and the higher survival of 1st and 2nd class.
* There was inconsistency between the importance of the other features highlighted by the Decision Trees (Age and Fare), as compared to one other feature highlighted by Lasso (Title\_Master). There is a natural link between “Age” and the Title “Master” (used for children), and there is also an evident link between “Fare” and “Class\_1” and “Class\_2”. This indicates that further analysis with different feature engineering may provide a more accurate model.
* Other techniques could be explored, like using the XGBoost (eXtreme Gradient Boosting).
* **IN THE END, IT HAS BEEN DETERMINED THAT SVM RBF RESULTED IN THE BEST MODEL WHEN TESTING AGAINST THE TEST DATA SET, FOLLOWED BY PRUNED DT, THEN RANDOM FOREST.**

SVM RBF IS THE WINNER!