Nguyen, Jenny

Final Paper

The goal of this project was to predict the survival of passengers on the Titanic using machine learning techniques. I implemented a logistical regression with an accuracy of 80% as my baseline to compare my other datasets to. Then a Nearest Neighbors (KNN) model, decision tree classifier, and bar graphs for visualization. I was able to classify passengers as either survivors or non-curvivors based on features such as age, sex, fare, and class. Understanding the survival patterns from the Titanic dataset is beneficial because it demonstrates how machine learning can be used for classification tasks, like historical datasets. Dealing with challenges of missing data and feature imbalances. This project is also an experimentation with exploring predictive analytics and handling real world data complexities.

The significance of this project is to predict survival rates on the Titanic. While it may sound unimportant, this dataset has broader implications in fields like risk analysis and decision making. Understanding patterns in survival situations, researchers and analysts can apply these similar methods to domains like health care, disaster management, and financial risk prediction. Also, through this project and running into problems, it demonstrates the importance of feature selection and data preprocessing, which are critical throughout this course and for the real world machine learning applications.

The dataset I used in this project is the Titanic dataset found from Kaggle. It contains 891 observations and 12 features. Qualitative variables include the Gender, Embarked, and Named. Quantitative variables include the Age, Fare, Class, and number of family members. The target variable is a binary with 0 being not survived and 1 being survived. This dataset was suitable for this project because of the larger number of observations and 12 features that are useful for classification problems and also providing challenges when running into missing values and imbalanced features.

The initial exploration reveals significant patterns in survival rates across the features. For example, when looking at the women and children in first-class they had a higher survival rate compared to Men. The summary statistics highlight the mean, median, and variance for numerical features such as Age and Fare. The qualitative features such as Embarked and Gener show a strong correlation with survival. When first working with the data, there was quite a bit of missing and unbalanced data. To address these issues I handled them by imputing median for the missing value of ages in my decision tree, which I realized now that I should’ve also done the same thing for my KNN and Logistic Regression, rather than just dropping the columns that had the missing values to keep it consistent. However, from my Logistical Regression a 80% accuracy score and 82% f-1 score for did not survive and 72% for survived isn’t bad. However, with my KNN there could be an improvement, especially with an accuracy of 68%. I also dropped features such as Embarked, which stated earlier was an important feature to test survivability, however, processing the data and turning each location embarked from caused my code my error and in the end I ended up not using it. I also dropped columns like “Cabin” because there was an excessive amount of missing data, this would’ve caused too many outliers and an unwanted low accuracy score. I then applied normalization to quantitative features such as “Fare” due to it being an unimportantly large integers, this makes sure the KNN model performs effectively. I also used mode imputation for categorical variables such as the “Embarked” Visualizations I used including bar graphs to illustrate who survived and didn’t survive based on gender, which was what I mainly wanted to look at due to historical context. In the future it would be more interesting to separate the age, so we can get a bigger understanding of gender, age (young, mid, old) and see where the demographics of who is more likely to survive based on these features. I also added a visualization from the features I had chosen what affected the survivability rate more. From that bar chart we’re able to see that age is an important feature, then gender, and then fare. I added features such as number of siblings, class, and parch to see if those features provide any significance, and turns out not as comparable as the other features when it comes to survivability. I experimented with the KNN visualizations as we can see Gender vs Age, Gender vs Fare, Class vs Fare, and Gender vs Class all have one thing in common, my mistake of comparing binary to continuous data. This resulted in having vertical clustered lines that became a headache to read. However, I did more experimenting and found that Age vs Fare shows those who are more likely to survive. Which is interesting to see how the fare numbers and Age intertwine with each other for survivability rate. From the KNN model we’re able to see a cluster of yellow (survived) to be clustered at the bottom left corner (younger age) of those that survived, and also quite a bit ranging between 25-35.

Models I decided to choose to approach this classification task were KNN, Decision Tree Classifier, and Logistic Regression. Logistic Regression was selected as a baseline and as a parametric method to evaluate the linear relationships between features and those that survived/didn’t. I also used it as a baseline to compare the other two datasets to. With an accuracy of 80% which just means that it correctly classifies 80% of the data points in the dataset, which isn’t too bad. Decision tree was used because it can uncover relationships between the features and the survival outcomes. However there should’ve been more implementation in my Decision trees such as including visualizations to see how the decisions were made, but unfortunately I only got to an accuracy score of 72% which is average. Playing around with the Decision trees would provide more insight on the features vs survived. KNN was selected to classify the passenger's survival status based on survival outcomes of the similar passengers in the dataset. It’s also important that KNN makes minimal assumptions and how it handles nonlinear relationships without requiring a lot of manual feature engineering. It was also very adaptable to missing data, which made preprocessing the data easier. However, I was scared of under or overfitting which happened a little in my data due to the choice of k value. When implementing these models for Logistic Regression I first standardized the continuous features, changed the column of gender to 0 and 1 for male and female for better tuning. Then I evaluated feature coefficients to determine what the most influential predictors of survival were. For the KNN model I normalized the continuous features such as age and fare to make sure there was equal contribution to the distance metric. Comparing accuracy score was important for KNN as it was optimal for hyperparameter tuning k. Choosing a k value is important because it affects the accuracy score and how well the model will perform and predict our features vs target. Errors I’ve found within my KNN model were working with continuous vs binary features, this resulted in a disproportionate influence of outlier, and a weird vertical line that shouldn’t be used for a dataset, this is due to the distance metric, when you only have values like 0 and 1 compared to 1-1000 it can be conflicting in the dataset and scaling the features the same way is important when it comes to modeling and predictions especially shown in this KNN model. However, there wasn’t any way to standardize gender vs all the other features due to it being binary. For the Decision Tree Model I used Gini to select optimal splits, then pruned the tree to prevent any overfitting by setting a minimum sample size for the splits. The accuracy score for the Decision tree was 72% which is just average. However, I should’ve provided more visualizations within my dataset so we can see how each decision was made based on features vs survivability. Logistic Regression played a critical role in modeling the linear relationships between the features and survival. The coefficients showed how there was a significant positive correlation of being female on the Titanic's survivability. KNN provided an important insight of how different features, tuning the K values contribute to the classification and how choosing binary vs continuous posted a problem.

The results from my models include, KNN an accuracy score of 68%, precision of 64%, and a F-1 score of 63% all not too bad but there is room for improvement such as adjusting my parameters and adjusting the k value to maximize the accuracy, precision, and F-1 score. FInding the correct K value affects the accuracy, and can also lead to over or underfitting. From my KNN model it accurately predicts survival outcomes for 68% of the passengers but 32% were misclassified. This could also be due to the metrics shown in my KNN models, choosing binary vs continuous could play into the effect of this issue. For my Decision tree there was an accuracy score of 72% which means that within my decision tree there were 72% decisions that were predicted correctly, and 28% that weren’t. In the future it’s important that I check this decision tree with either Confusion Matrix to examine the number of true positives, true negatives, false positives and false negatives. Cross Validation to check if the score is consistent across multiple test splits. Finally, from my Logistic Regression there was an Accuracy of 80%, Precision of 77%, and a F-1 Score of 75%. 80% of the time my model accurately predicts if a passenger has survived or not survived. Precision is 77% in passengers that did survive, and F-1 Score means there's a balance between my precision and recall. In the future I can try to improve my Logistic Regression by adjusting the thresholds for classifying a passenger as “survived” to optimize the precision. Adjusting my features can improve my model performance, and trying other algorithms such as cross-validation, regularization, and hyperparameter tuning.

In Conclusion working with this dataset was fun, however, there are a lot more things I can do to improve upon, such as providing more algorithms or adjusting my features more, and even creating more visualizations because accuracy score itself isn’t enough such as in my Decision Tree. I’ve understood the importance of data preprocessing across all my models, especially for my KNN and adjusting the parameters and choosing the right values plays an important role for how my model will perform. Having other models to check my algorithms such as confusion matrix and cross validation are very important to seeing what features may be affecting my dataset. Overall, it was enjoyable to work with this classification model, but there need to be more additional things to make my model more accurate to predict the survivability. However, my hypothesis was correct that more women were likely to survive than men. It would be interesting though if I had separated the age into different categories and looked at what the survival rate was in a bar chart. But we were able to capture it in the KNN, and saw that younger and 25-35 people of that age were more likely to survive. It would also be interesting to see it based on gender vs those ages too, but as we saw classification and binary didn’t work out too well. From these models and this final project I was able to highlight the importance of different ways I can approach my classification task, implement it, see my errors, and what I can improve upon.

In my future career I hope to be able to understand the complexities of working with large datasets for data analysis, as I want to get into program management. Overlooking and managing a lot of projects is important, and understanding datasets within a company is beneficial in growing and gaining profit within a business. I hope to be able to use these skills to improve upon but also learn how to implement models from scratch and enhance my programming proficiency. From this I was also able to understand trade-offs in model evaluation and selection. From this project I have gained more experience in working with real-world datasets and the importance it is to have the right models and evaluation.

Resources;

Videos that helped me build these models from scratch.

Decision trees:

[How to implement Decision Trees from scratch with Python](https://www.youtube.com/watch?v=NxEHSAfFlK8)

KNN:

[How to implement KNN from scratch with Python](https://www.youtube.com/watch?v=rTEtEy5o3X0)

Logistic Regression:

[Logistic Regression from Scratch - Machine Learning Python](https://www.youtube.com/watch?v=x1ez9vi611I)

Resources that helped me understand my models:

Cuello, Fransisco. “K-Nearest Neighbors (KNN) | TrendSpider Learning Center.” *Trendspider.com*, 8 Aug. 2024, [trendspider.com/learning-center/k-nearest-neighbors-knn/](http://trendspider.com/learning-center/k-nearest-neighbors-knn/).

GeeksForGeeks. “Decision Tree Introduction with Example.” *GeeksforGeeks*, 16 Dec. 2017, [www.geeksforgeeks.org/decision-tree-introduction-example/](http://www.geeksforgeeks.org/decision-tree-introduction-example/).

GeeksforGeeks. “Understanding Logistic Regression.” *GeeksforGeeks*, 9 May 2024, [www.geeksforgeeks.org/understanding-logistic-regression/](http://www.geeksforgeeks.org/understanding-logistic-regression/).

Rajib Kumar Halder, et al. “Enhancing K-Nearest Neighbor Algorithm: A Comprehensive Review and Performance Analysis of Modifications.” *Journal of Big Data*, vol. 11, no. 1, 11 Aug. 2024,<https://doi.org/10.1186/s40537-024-00973-y>.

‌

Historical:

Tikkanen, Amy. “Titanic | History, Sinking, Rescue, Survivors, & Facts.” *Encyclopedia Britannica*, 26 Oct. 2018, [www.britannica.com/topic/Titanic](http://www.britannica.com/topic/Titanic).

“Data-Driven Survival: Titanic Edition.” *Www.shiftcomm.com*, [www.shiftcomm.com/thinking/never-let-go-titanic-survival-101](http://www.shiftcomm.com/thinking/never-let-go-titanic-survival-101).

‌

‌