TITANIC - MACHINE LEARNING FROM DISASTER

- **Team 7 - Members:** Hayes Justin Landry and Jakob West
- **Course Information:** CS3820-002
- **Professor:** Dr. Moin
- **TA:** Himon Thakur

TOPIC: TITANIC SURVIVAL PREDICTION

- **Customer and End-User:** Kaggle, data scientists; historians
- **Problem Statement:** Build a Machine Learning model and compete in a Data Competition to predict who survived vs who did not survive the Titanic using historical data.
- **Value Proposition:** Contributes to work done on Kaggle; adds to leaderboard; to help the data science community (Data science enthusiasts and students)

RELATED WORK AND UNIQUE SELLING POINT

- **Related Work:** Email, Cancer screenings, Stock Markets, Sports
- **Unique Selling Point (USP):** Individual contributions
- **Reuse of Existing Work:** experimenting with ideas from other Kaggle contributors

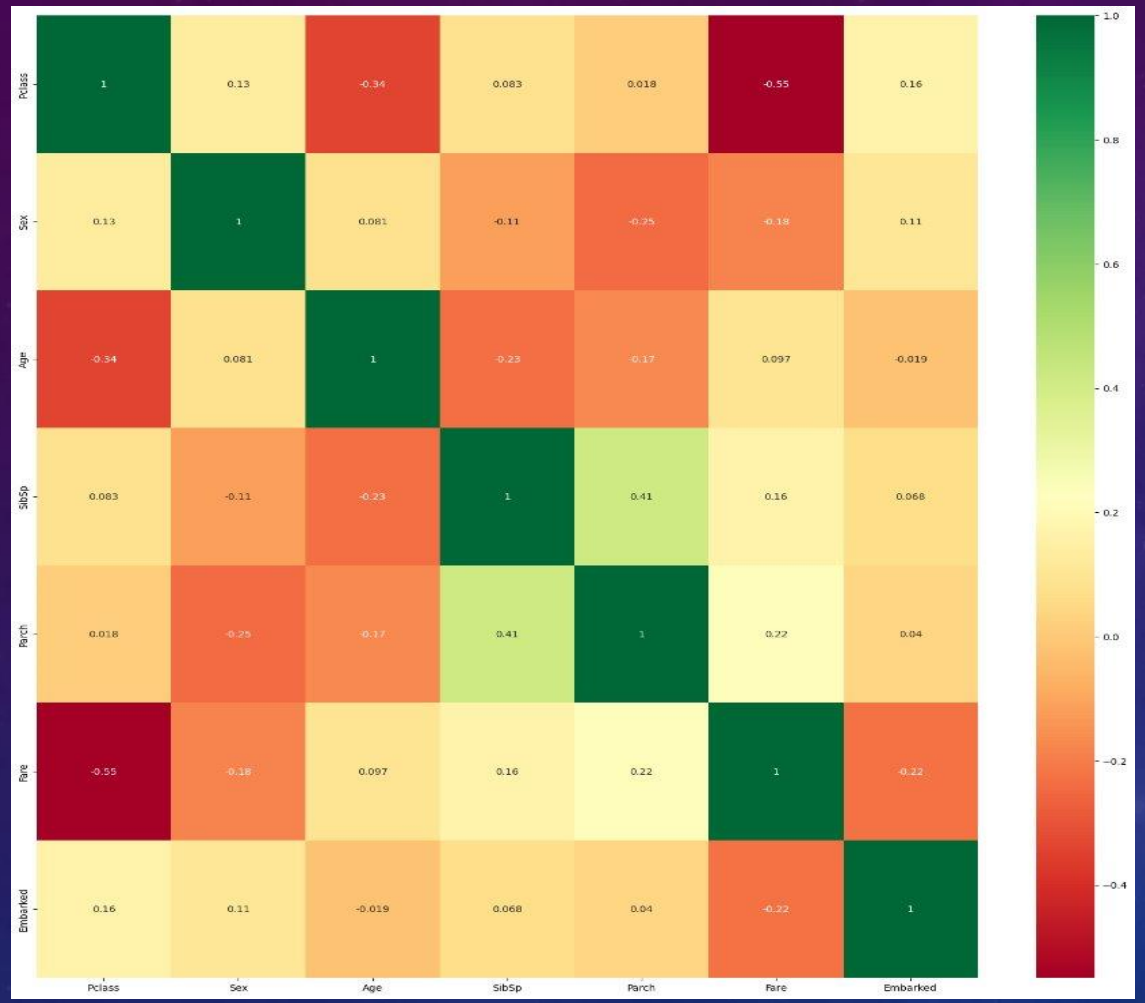
WORK BREAKDOWN STRUCTURE

- Individual Tasks
  - Data Understanding & Preprocessing
  - Feature Engineering (Justin)
  - Model Selection & Training (Jakob)
- Group Tasks
  - Evaluation and Reporting

OPEN-SOURCE CONSIDERATION

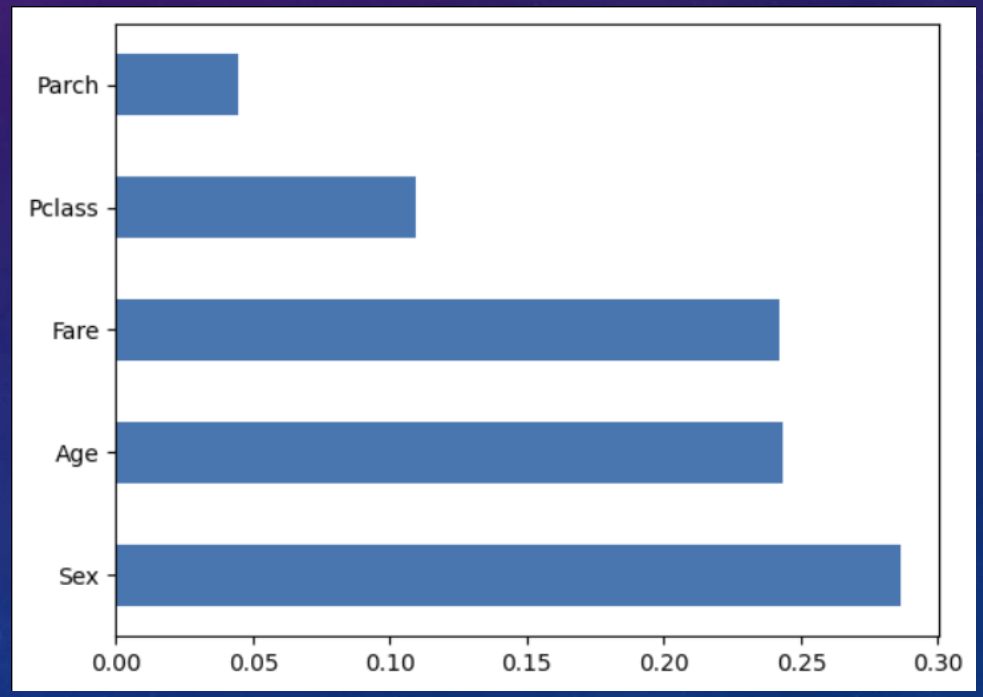
- We are looking into open source depending on rules set by Kaggle.com
- MIT license is our first choice

COURSE CONTENT RELATIONSHIP (1 / 2)



Initial training scores

```
Decision Tree Accuracy: 0.8045
Random Forest Accuracy: 0.8380
Logistic Regression Accuracy: 0.8101
Gradient Descent Accuracy: 0.7318
XGBoost Accuracy: 0.8547
```



COURSE CONTENT RELATIONSHIP (2 / 2)

- Binary classifier
- Scikitlearn’s impute and preprocessing libraries
- **Models for consideration**: Naive Bayes, Logistic Regression, K-Nearest Neighbours, Support Vector Machine, Decision Tree, Random Forest ,Voting Classification, XGBoost, Gradient Descent, Deep Learning with a neuronal network
- **Feature Engineering:**
  - Justin: categorizing Age, combining sibsp and parch categorizing into different size families, look into fare and pclass
- **Model Evaluation Metrics:**
  - Jakob: determining which model(s) to commit to

Data Dictionary

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

KAGGLE COURSES

Completed

Intro to Programming

Next up: Exercise: Arithmetic and Variables

Python

Next up: Exercise: Syntax, Variables, and Numbers

Intro to Machine Learning

Next up: Exercise: Explore Your Data

Upcoming

Intermediate Machine Learning

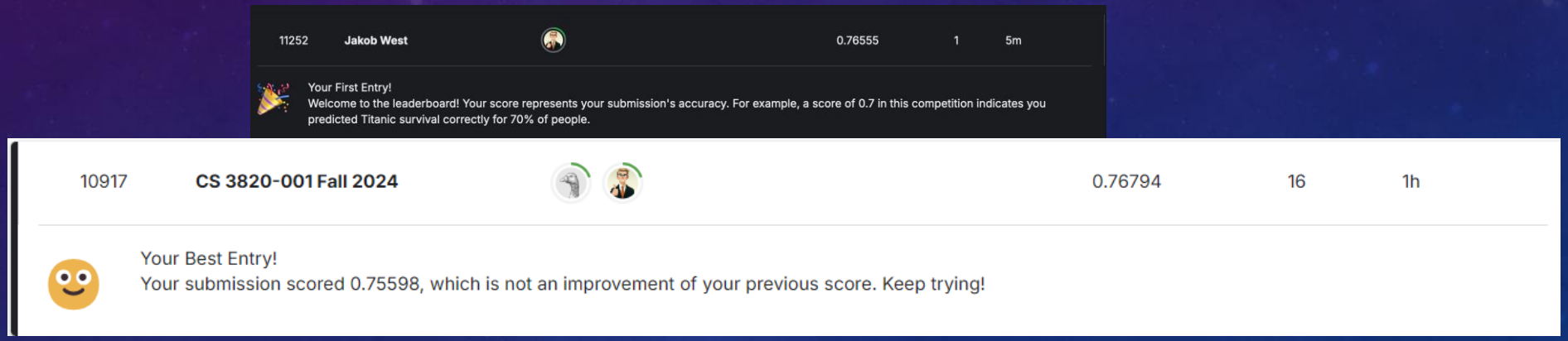
Next up: Exercise: Introduction

Feature Engineering

Next up: Mutual Information

PROJECT SUMMARY & CONCLUSION

- Current accuracy score of .76794
- Goal accuracy score of .80, which would place us into the top 800 on Kaggle.com



TIMELINE

- Week - 10/28 - 11/03 --> Monday 10/28, half-way presentation; Feature Engineering Kaggle Course
- Week - 11/04 - 11/10 --> Apply feature engineering techniques; optimizing correlation matrix
- Week - 11/11 - 11/17 --> Intermediate Machine Learning Kaggle Course; application of course
- Week - 11/18 - 11/24 --> evaluation / performance metrics; finalize models and submissions
- Week - 11/25 - 12/01 --> Thanksgiving; time with family and break from school; so necessary
- Week - 12/02 - 12/08 --> Wednesday, 12/04, final presentation; turn in final deliverable

QUESTIONS?

- References:
  - Kaggle. (n.d.). Titanic: Machine Learning from Disaster. Retrieved from https://www.kaggle.com/c/titanic