Titanic Survival

Nicholas Morris

Machine Learning

Wrangling

Text Manipulation

Extracting the title from passenger's names. Extracting the first character from ticket and cabin. Filling in missing values with k-Nearest Neighbors.

Feature Engineering

Binary Data

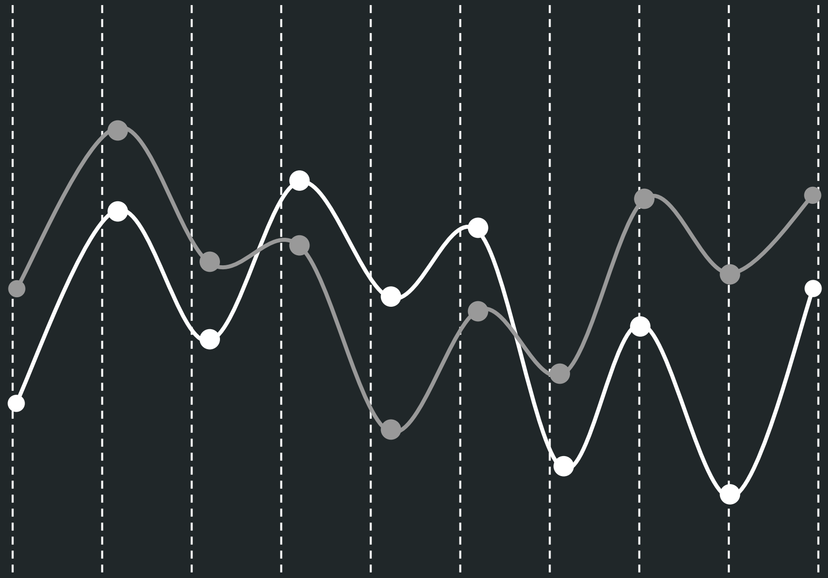
Converting Pclass, Title, Sex, SibSp, Parch, Ticket, Cabin, and Embarked to binary data points.

Modeling

Predictions

Training logistic regression, XGBoost, and deep learning neural network models. Evaluating performance. Computing feature drift to signal retraining.

Wrangling



Dataset

Below is the first two passengers in the data. There are 891 total passengers. The target we are predicting is Survived. [Link to the dataset]

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C

Title

Survived	Title
0	Mr.
1	Mrs.
1	Miss.
1	Mrs.

The title in each passenger’s name was extracted.

Ticket

Survived	Ticket
0	A
1	P
1	S
1	1

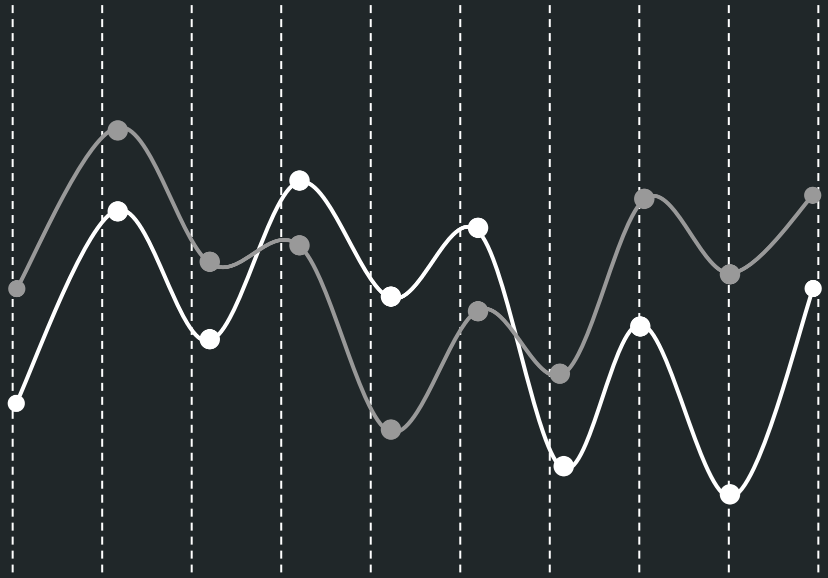
The first character in Ticket was extracted.

Cabin

Survived	Cabin
0	N
1	C
1	N
1	C

The first character in Cabin was extracted.

Feature Engineering



Binary Data

Pclass, Title, Sex, SibSp, Parch, Ticket, Cabin, and Embarked were converted to binary variables.

Pclass_1	…	Sex_male	SibSp_1	…	Parch_0	…	Ticket_A	…	Cabin_N	…	Embarked_S	…	Title_Miss.	…	Survived
0	…	1	1	…	1	…	1	…	1	…	1	…	0	…	0
1	…	0	1	…	1	…	0	…	0	…	0	…	0	…	1
0	…	0	0	…	1	…	0	…	1	…	1	…	1	…	1
1	…	0	1	…	1	…	0	…	0	…	1	…	0	…	1

Atwood Numbers

An Atwood Number is a calculation that shows the relative change between two variables. The formula for two variables x and y is:
$(x - y) / (x + y)$

This calculation was done on all pairs of non-binary variables; but did not improve model performance, so, it was left out of the final model.

Binning

Binning is when a non-binary variable is grouped into histogram bins, and represented as binary variables.

Binning did not improve model performance, so, it was left out of the final model.

Reciprocals

A reciprocal is when a non-binary variable x is calculated as 1 / x.

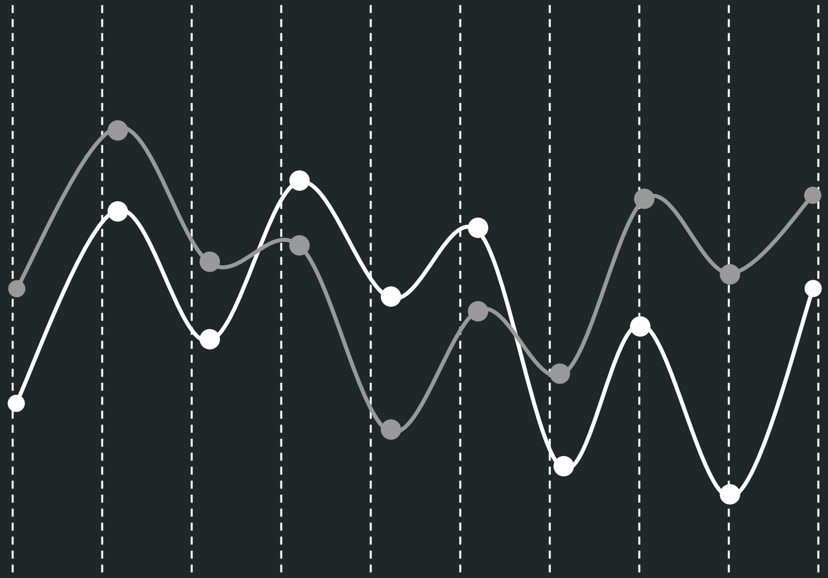
Reciprocals did not improve model performance, so, it was left out of the final model.

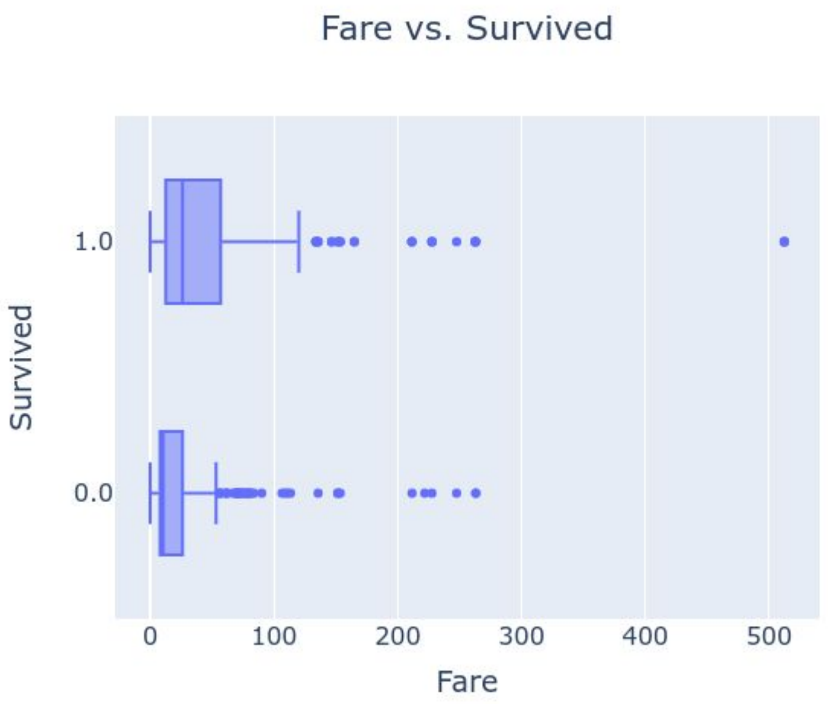
Interactions

An interaction is when two variables x and y are calculated as x * y. Reciprocals were fed into this calculation to generate x / y as well.

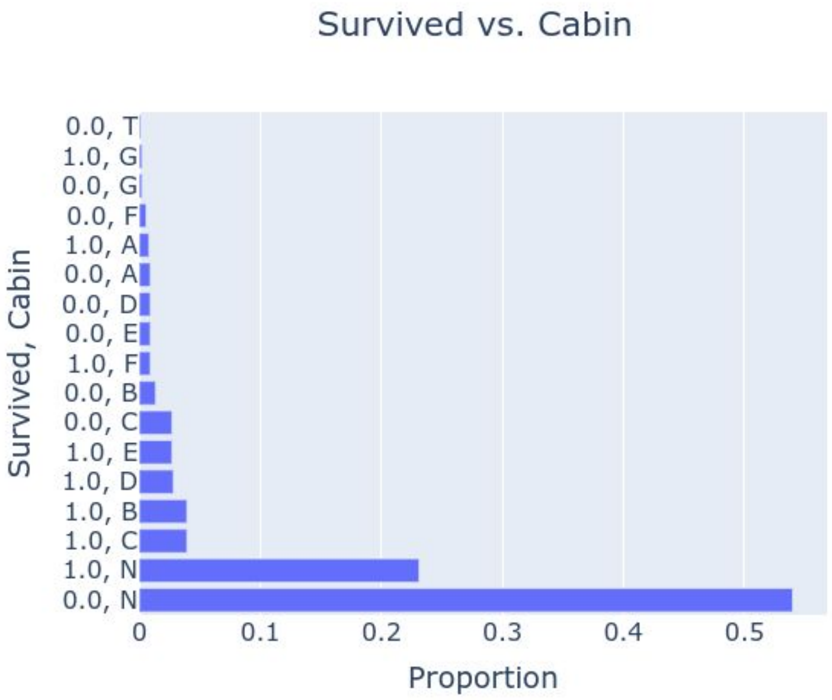
Interactions did not improve model performance, so, it was left out of the final model.

Data Exploration

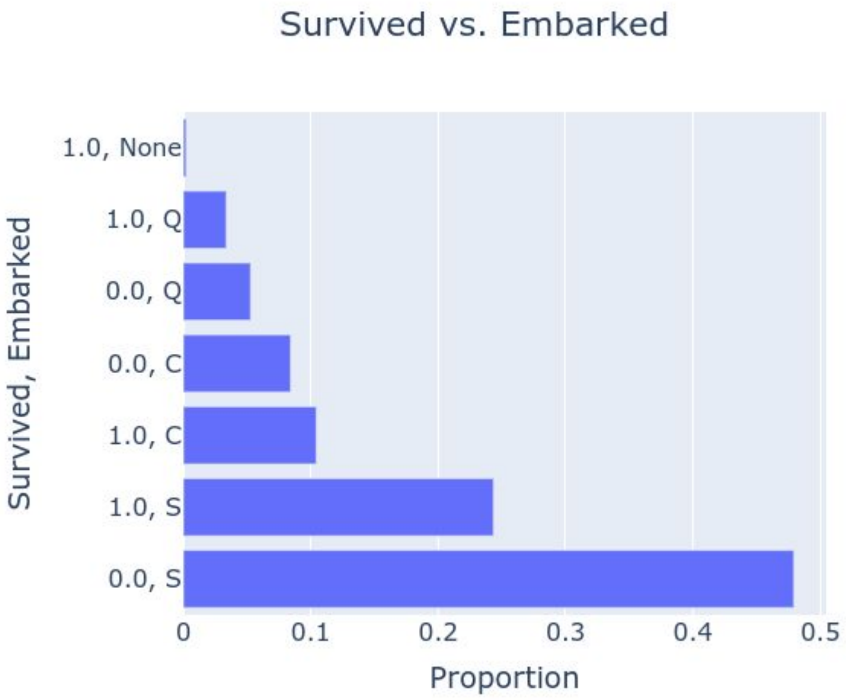




We can see that passengers who paid a higher fair are slightly more likely to survive. Survivors paid a median fare of $26. Deceased paid a median fare of $10.5.

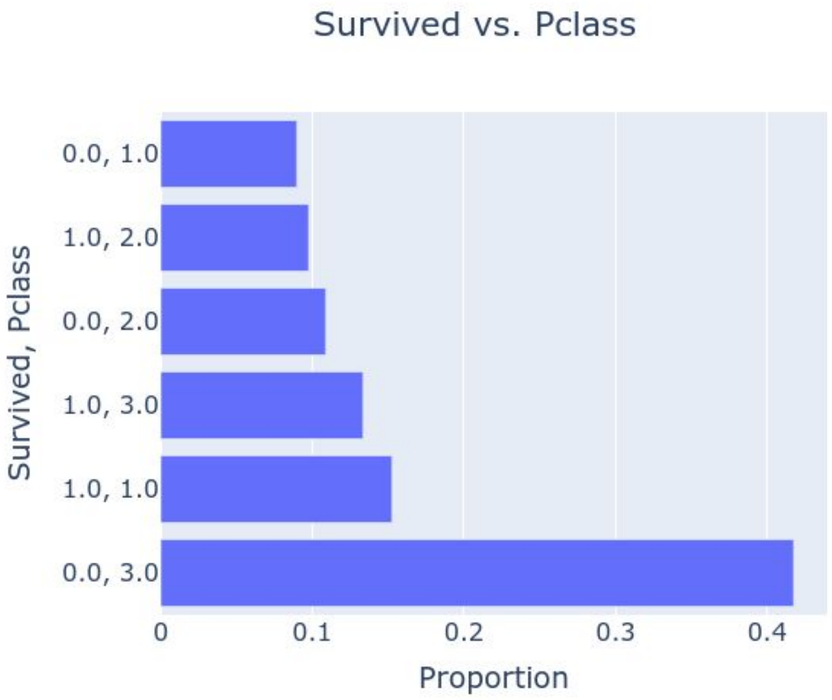


We can see that passengers who didn't have a cabin (N) were much more likely not to survive. 54% of the passengers were deceased and without a cabin.

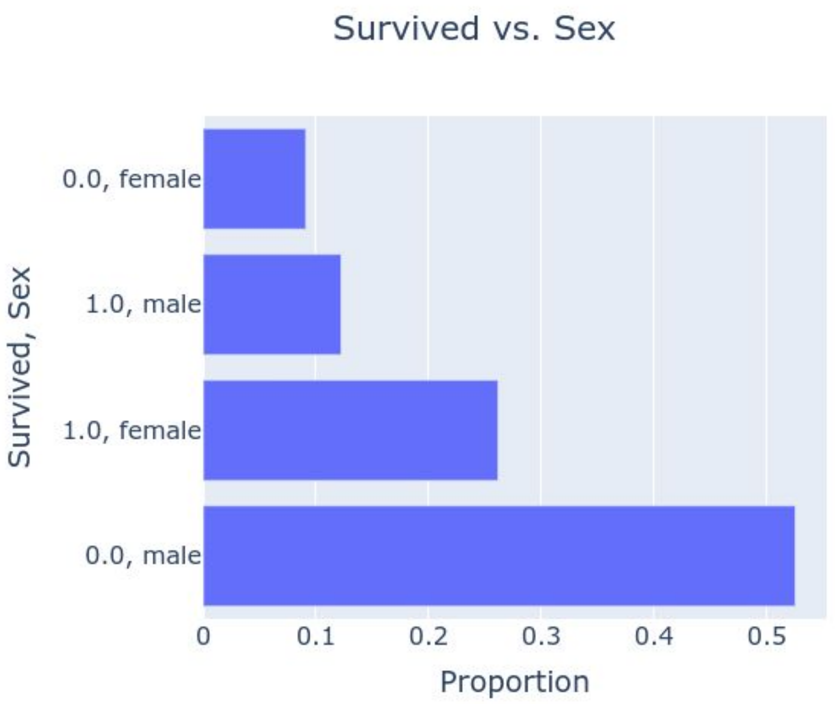


We can see that passengers who embarked from port S were much more likely not to survive. 48% of the passengers were deceased and from port S.

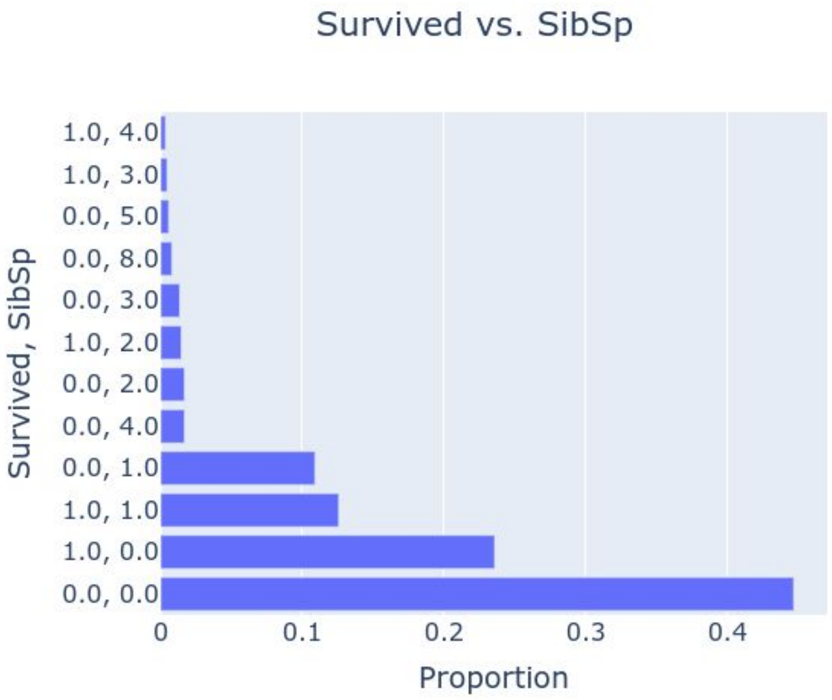
We can see that passengers who had no parents on board (Parch = 0) were much more likely not to survive. 50% of the passengers were deceased and had no parents on board.



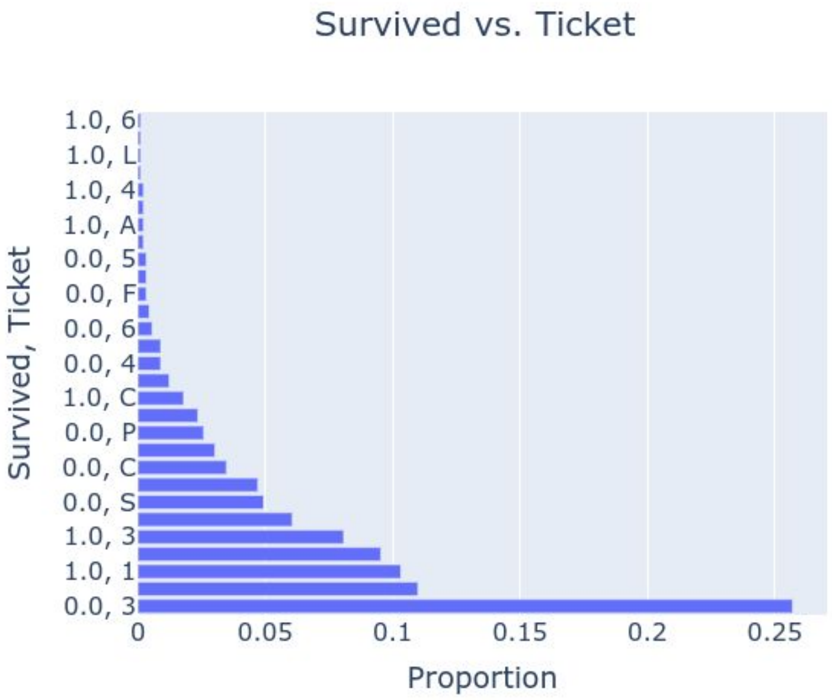
We can see that passengers who had the lowest class ticket (Pclass = 3) were much more likely not to survive. 42% of the passengers were deceased and had the lowest class ticket.



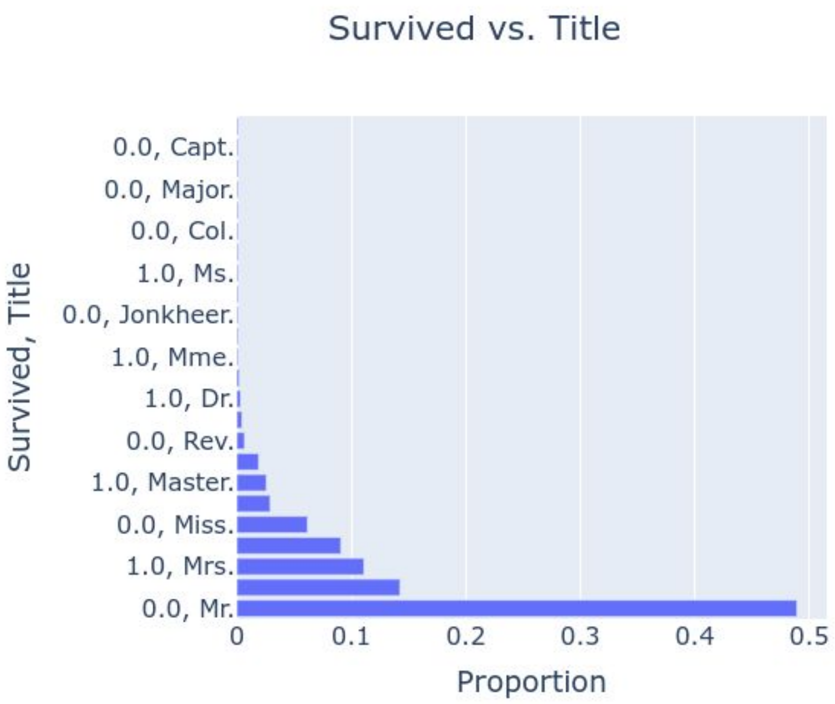
We can see that male passengers were much more likely not to survive. 52% of the passengers were deceased and male. We can also see that female passengers were much more likely to survive. 26% of the passengers survived and were female.



We can see that passengers who had no siblings on board (SibSp = 0) were much more likely not to survive. 44% of the passengers were deceased and had no siblings on board.

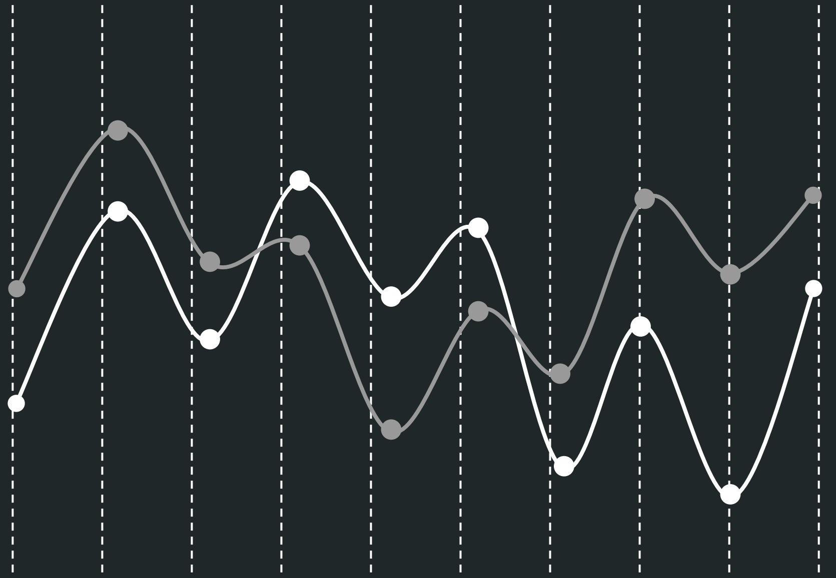


We can see that passengers who had a ticket number starting with 3 were much more likely not to survive. 26% of the passengers were deceased and had a ticket number starting with 3.



We can see that passengers who had a title Mr. were much more likely not to survive. 49% of the passengers were deceased and had a title Mr.

Modeling



Model Parameters

Logistic Regression

Library: scikit-learn
Penalty: L1
Number Of Alphas: 16
Cross Validation Folds: 3
Tolerance: 1e-4
Max Iterations: 100

XGBoost

Library: xgboost
Boosting Rounds: 100
Learning Rate: 0.1
Max Depth: 7
Min Child Weight: 1
Column Sampling: 0.8
Row Sampling: 0.8

Neural Network

Library: Tensorflow
Epochs: 500
Learning Rate: 0.001
Batch Size: 16
Layers: 10
Nodes Per Layer: 128
Solver: Adam

Model Comparison

Logistic Regression

Accuracy: 0.86
F1: 0.85
In Control: 100%

Model Indicators:
1. Title_Master.
2. SibSp_5
3. Parch_4
4. SibSp_8
5. Sex_male

XGBoost

Accuracy: 0.82
F1: 0.79
In Control: 100%

Model Indicators:
1. Sex_male
2. Title_Rev.
3. Title_Mr.
4. Pclass_3
5. Sex_female

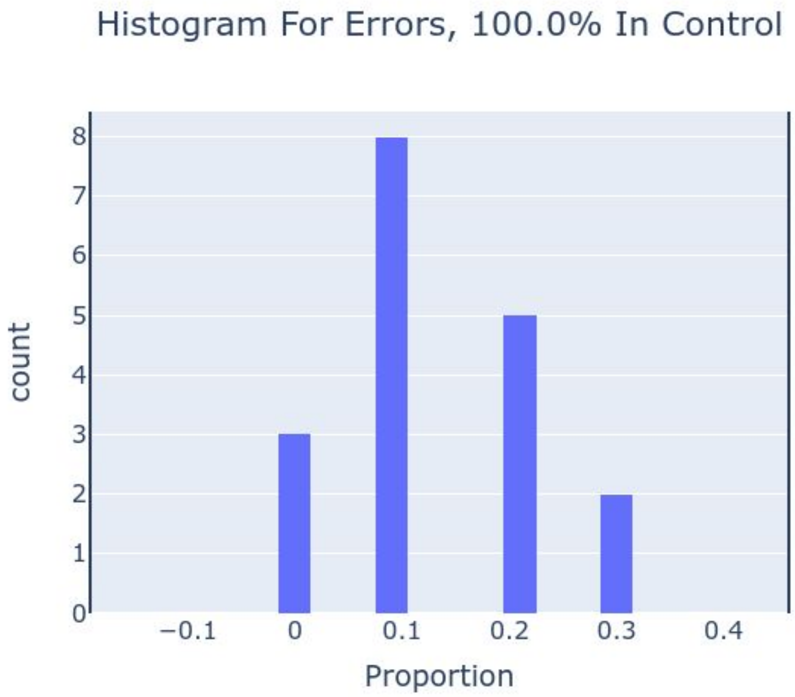
Neural Network

Accuracy: 0.78
F1: 0.73
In Control: 100%

Model Indicators:
1. Parch_6
2. Parch_5
3. Parch_4
4. Parch_3
5. Parch_2

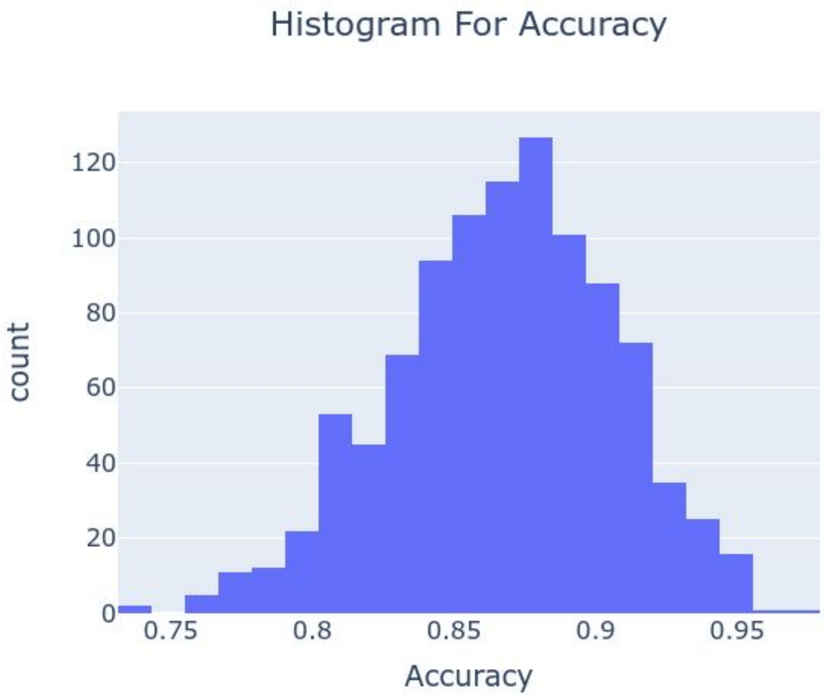
	Predicted Deceased	Predicted Survived
Deceased	48.8%	5.6%
Survived	7.8%	37.6%

These predictions come from the logistic regression model. These predictions are done on 20% of the data that the model did not see during training. Only 13.4% (5.6% + 7.8%) of the predictions are wrong, and 86.6% of the predictions are correct. There's a slightly stronger tendency to predict survivors as deceased compared to predicting deceased as survived.



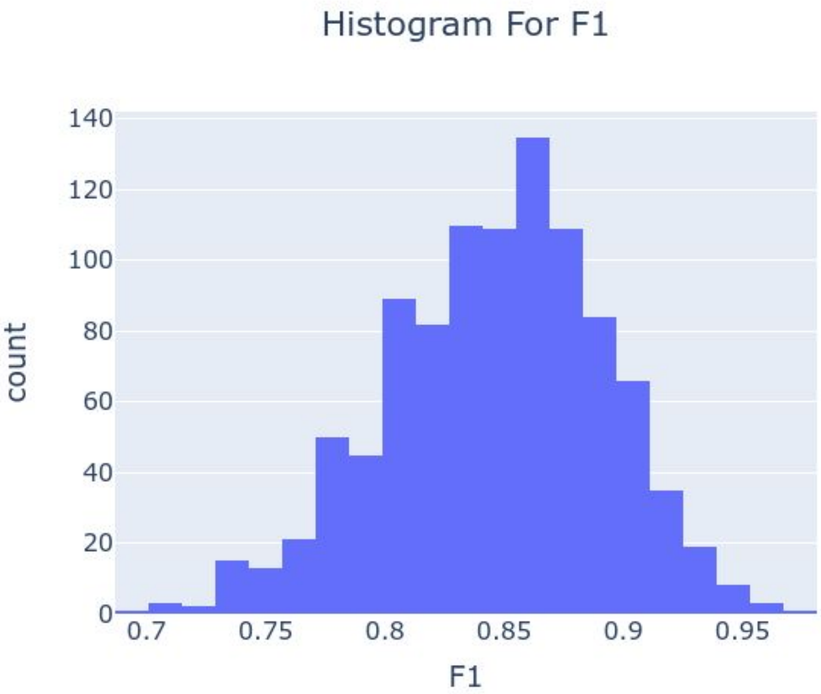
The errors are the fraction of 10 predictions that were wrong.

The errors are most likely to be 10%. Control limits were computed on the errors and we can see that the prediction error is completely under control.



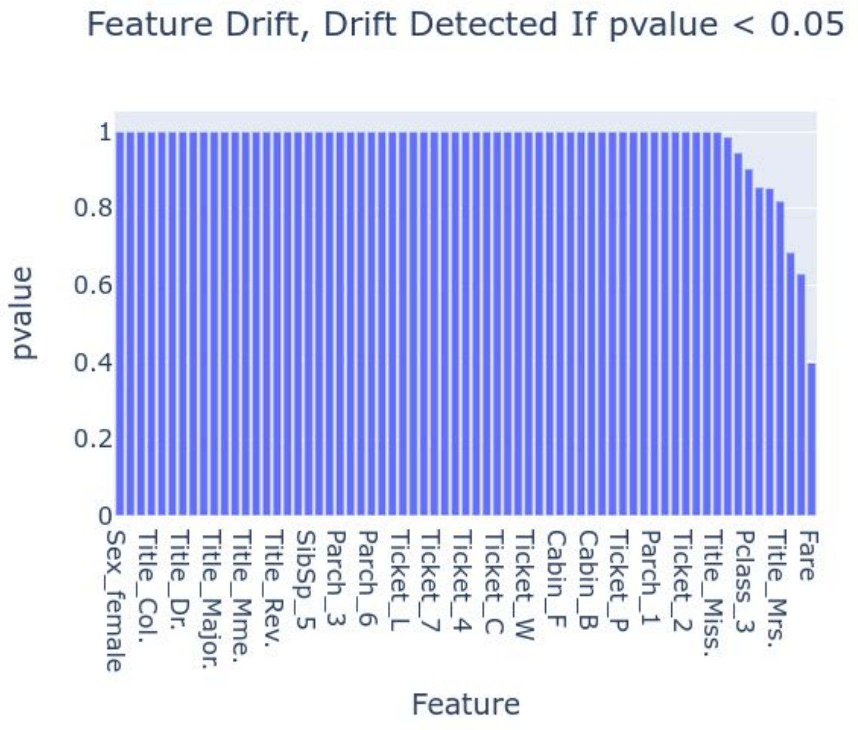
The prediction error was resampled 1000 times at a 50% sampling rate with replacement. Then Accuracy was computed on each sample to get a distribution.

Accuracy has a wide range between 0.75 and 0.95, which isn't good. Accuracy has a bell shape, which is good, and a slight skew to the left.



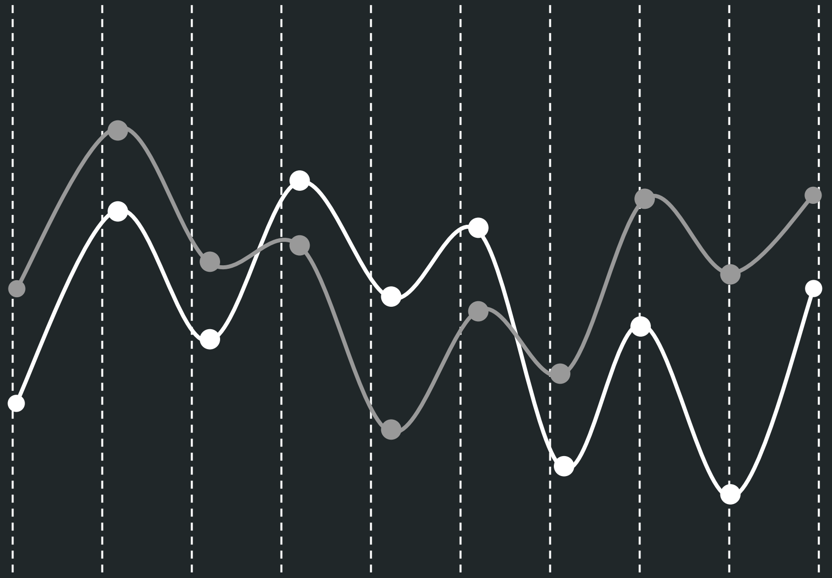
The prediction error was resampled as previously mentioned to get a distribution for F1. F1 is a combination of Precision and Recall. Precision tells us how well the model doesn't label the deceased as survived. Recall tells us how well the model finds all survivors.

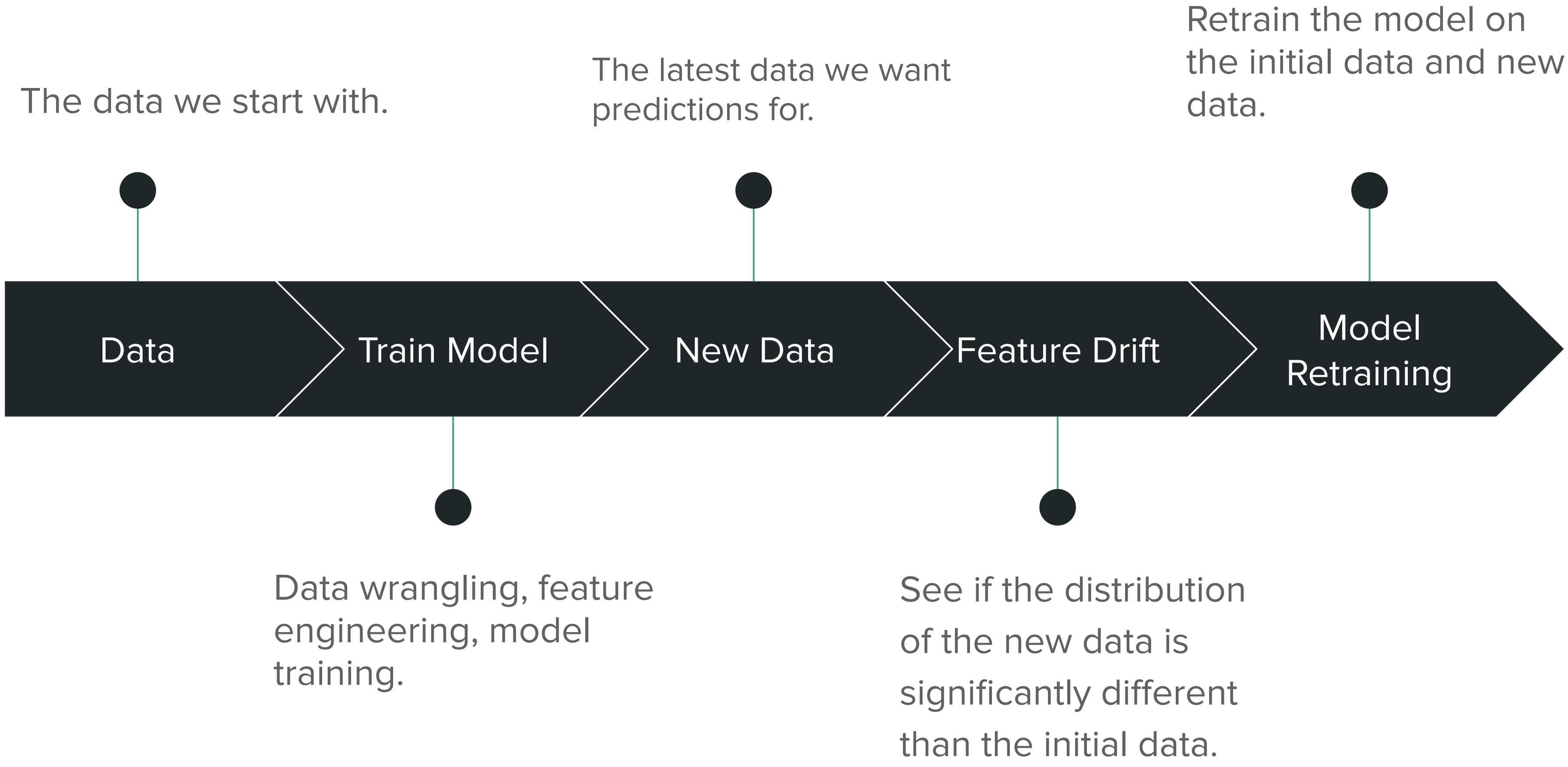
F1 has a wide range between 0.7 and 0.95, which isn't good.



A Kolmogorov-Smirnov test was performed for each column in the data to see if the distribution of the testing data is the same as the training data. If the testing data does not share the same distribution as the training data, then there is a drift, which signals for model retraining. All of the columns do not experience a drift, which is good.

Deployment




The data we start with.
The latest data we want predictions for.
Retrain the model on the initial data and new data.
Data
Train Model
New Data
Feature Drift
Model Retraining
Data wrangling, feature engineering, model training.
See if the distribution of the new data is significantly different than the initial data.

Thank You