Project 4 - Alphabet Soup

—

Hosted by Julia Begley, Hayley Fuller, Alemseghed Ghebrezghi, Diego Lorenzo, Lewis Russell

Project overview

Focus:

- Tableau shows us that around 75-80% of campaigns were classified as successful, so how can machine learning make this % better?
- Using Tableau what further analysis can we find. How do these findings compare to machine learning?
- What will machine learning do to predict a successful campaign?
- Can machine learning predict a higher success rate? If so then what does a successful campaigns need?
- Is the success rate prediction in machine learning the same as the data in Tableau?
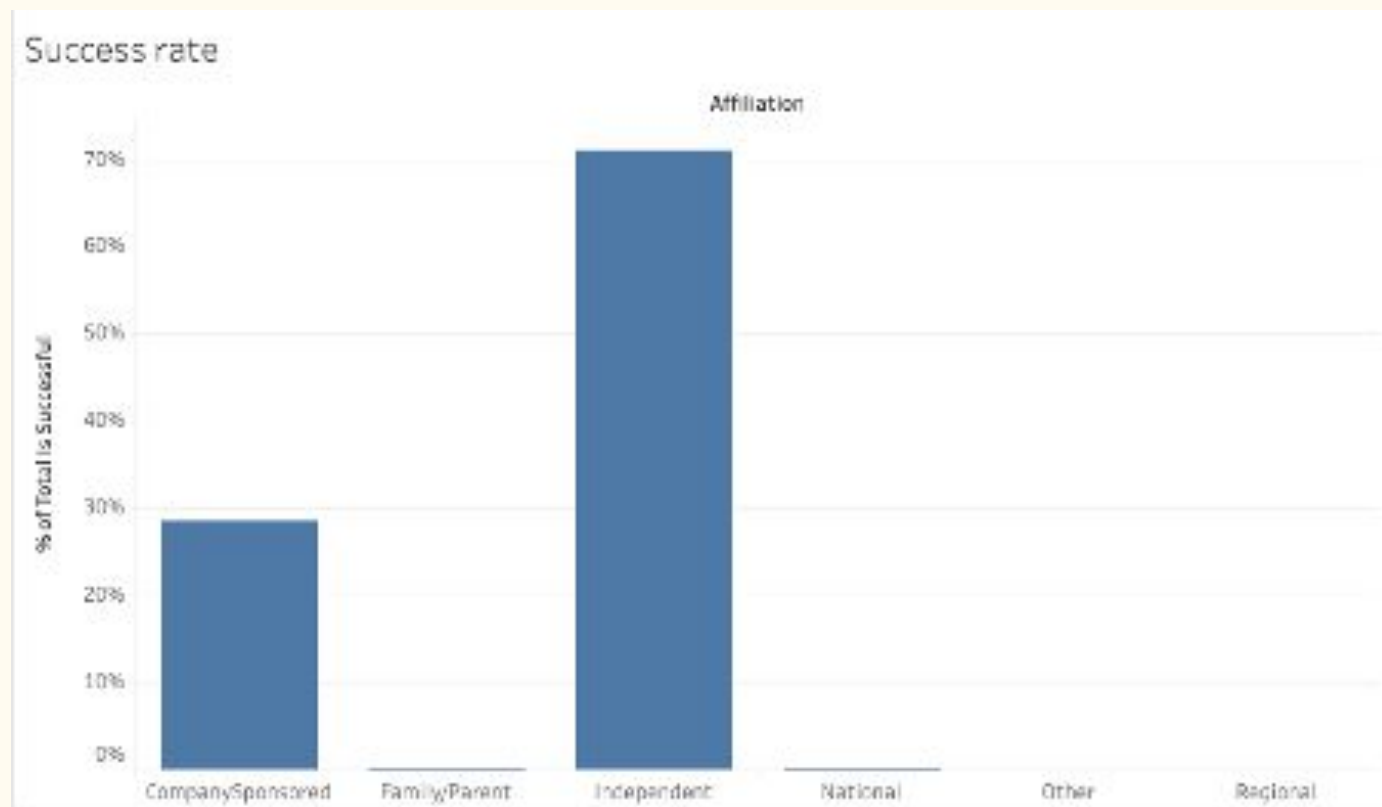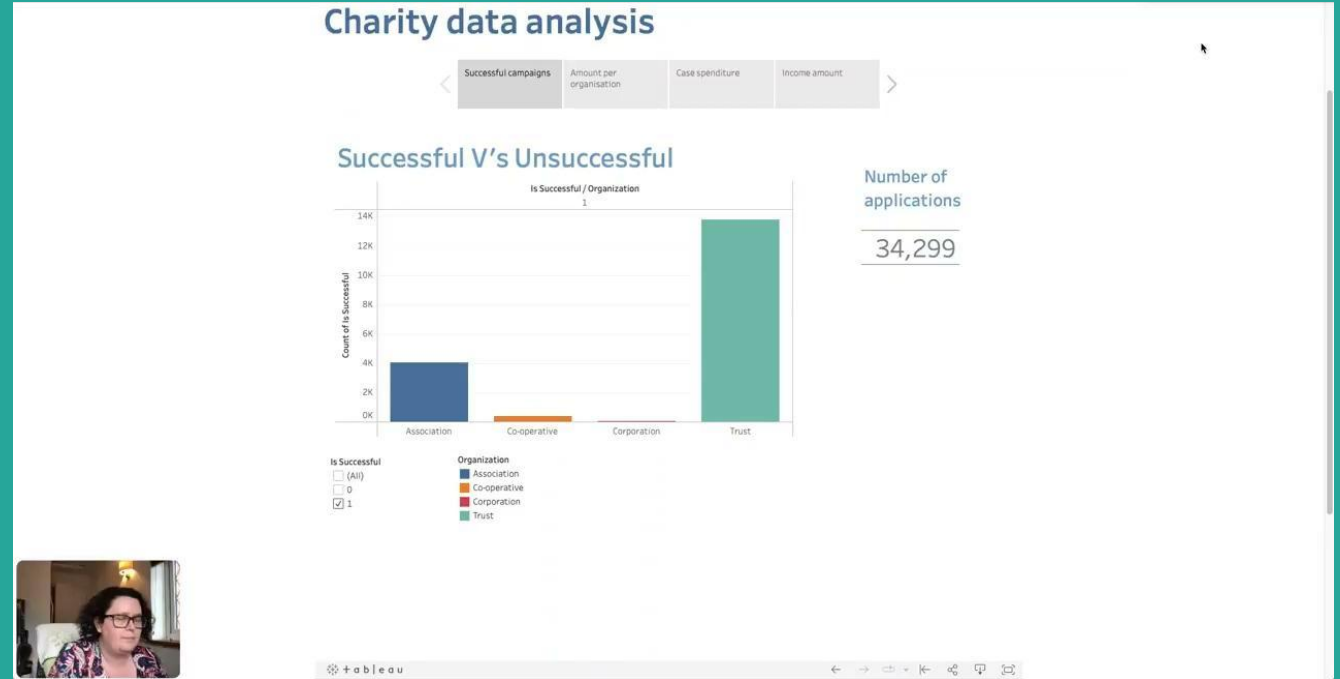
Tableau Analysis

Tableau Analysis

<u>Story 1 Dashboard</u>

ML Base code

1. Overview

- The Alphabet Soup Nonprofit Foundation sought a binary classifier in order to facilitate the selection of applicants for funding.
- The binary classifier is designed to assess the likelihood of successful applicants being funded by Alphabet Soup.
- Preprocessing, compiling, training, evaluating the model and model optimization have been done all in a single jupyter notebook file.

Project 4_ group 1: ML Base code

1. **Data processing**

- **Importing the csv file from API source**

    url = 'https://static.bc-edx.com/data/dl-1-2/m21/lms/starter/charity_data.csv'

    charity_data_df = pd.read_csv(url)

    charity_data_df.head()

- **Feature extraction - Removal of features : EIN and Name were eliminated from the dataset,**

- **Removal of redundant features for example outliers**

    # Drop the non-beneficial ID columns, 'EIN' and 'NAME'.

    Charity_data_df.drop(columns = ['EIN'], inplace=True)

    charity_data_df

- **Splitting**

➢ **Split our preprocessed data into our features and target arrays**

▪ **Spliting data set into features (independent variables) and target variable(dependent variable)**

    X = charity_data_dummies.drop('IS_SUCCESSFUL', axis=1).values

    y = charity_data_dummies['IS_SUCCESSFUL'].values

➢ **Split the preprocessed data into a training and testing dataset**

    X_train, X_test, y_train, y_test = train_test_split(X, y, random_state = 42)

- **Normalization or feature scaling**

Project 4_ group 1: ML Base code

2. Compiling, Training, and Evaluating the Model and optimize the model

– Three models(Attempts)

- Define the model - deep neural net, i.e., the number of input features and hidden nodes for each layer.

    input_features_total = len(X_train[0])

    hidden_nodes_layer1

    hidden_nodes_layer2

    hidden_nodes_layer3

    number_of_neurons = tf.keras.models.Sequential()

- Compile the model
- Tain the model
- Evaluate the model using the test data

    model_loss, model_accuracy = number_of_neurons.evaluate(X_test_scaled,y_test,verbose=2)

    print(f"Loss: {model_loss}, Accuracy: {model_accuracy}")

- Optimize the Model

Project 4_ group 1: ML Base code

3. Summary

Analysis result

- From the three optimised deep learning models,
- the accuracy has been found

    74, 57 and 76% for first, second and third attempts (models) respectively.

- The third attempt's result is just above the target value and hence it can be taken as a final optimum model.

Recommendation

For the model to continue to forecast with optimum accuracy,

- Increase number of hidden layers
- increase of number of nodes(neurons or preceptors)
- Removal of features that does not affect the result.

Model accuracy by attempt

These graphs show us that the models are working as expected because the training and validation lines both show accuracy increasing over time.
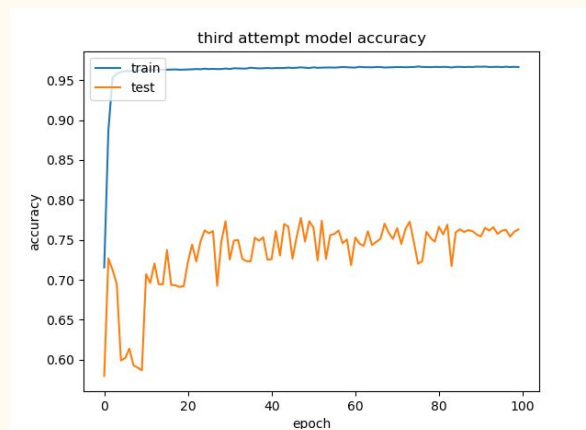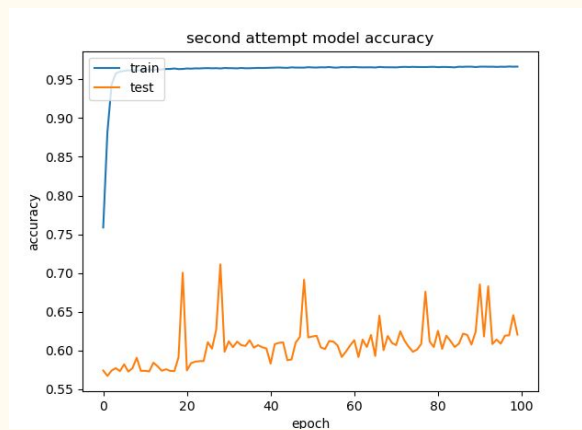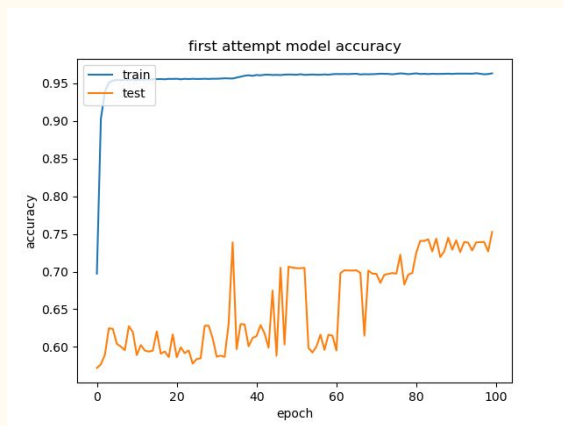
Model loss by model attempt

The loss graphs for the 1st and 2nd attempts at the model show they are working as expected as the training and validation loss scores both decrease over time.

However, the graph for the 3rd attempt indicates that it may be 'overfitting', i.e. matching too closely to the training data, and failing to fit well to the additional, testing data.
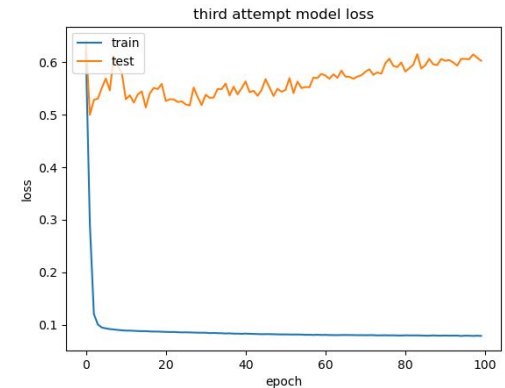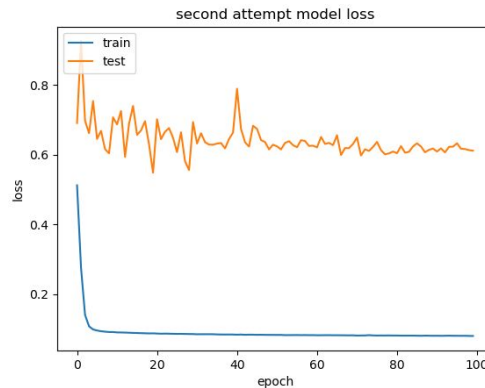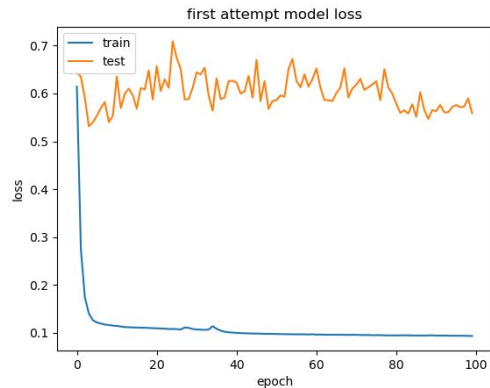
Classification Report for the test data

A Classification Report measures a model's quality of predictions using 3 metrics

	precision	recall	f1-score	support
0	0.52	0.93	0.67	4037
1	0.79	0.25	0.38	4538
accuracy			0.57	8575
macro avg	0.66	0.59	0.52	8575
weighted avg	0.67	0.57	0.52	8575

1. **Accuracy:** how often the model is correct, the percentage of correctly predicted observations to the total number of observations.
   The model was right/accurate **57%** of the time.

2. **Precision**: High precision relates to a low false positive rate.
   Out of the 34,000+ organisations that received funding from Alphabet Soup over the years, **79%** used the money effectively.

3. **Recall**: High recall relates to a low false negative rate.
   Of all the organisations that received funding the model correctly predicted **93%** of the time which ones didn't use the money effectively

Classification Report for the test data

	precision	recall	f1-score	support
0	0.52	0.93	0.67	4037
1	0.79	0.25	0.38	4538
accuracy			0.57	8575
macro avg	0.66	0.59	0.52	8575
weighted avg	0.67	0.57	0.52	8575

CONCLUSION

a)   Model was correct 57% of the time -not very accurate.

b)   Recall is the most reliable metric -model correctly predicts 93% of the time which organisations fail to use the money effectively.

In order to improve the model Accuracy it would be recommended to:

I.   **Add more datapoints**

II.   **Further Exploratory Data Analysis**:
E.g.: Avoid Class Imbalance (Oversampling vs Undersampling) in order to have a balanced dataset.

III.   **Feature Engineering:**
Selecting the features that best capture the pattern in the dataset by using tools like Random Forest or Gridsearch.

Conclusions

"We will then consider the differences, advantages and disadvantages of data analysis done by only humans vs. analysis supported/directed by Machine Learning"

In Tableau successful applications have

1. An income amount of between 25,000 to 99,999
2. Use the money to Preserve the charity
3. Are classified as a Trust
4. Be independently affiliated

In Machine Learning:

1. Model was correct 57% of the time -not very accurate.
2. Recall is the most reliable metric -model correctly predicts 93% of the time which organisations fail to use the money effectively.