

```
In [2]: # Import our dependencies
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, OneHotEncoder, MinMaxScaler
import pandas as pd
import tensorflow as tf
import numpy as np

# Import our input dataset
df = pd.read_csv('./pitcher_salaries_cleaned.csv')
df.head()
```

Out[2]:

	Year	Full Name	Age	Salary	ERA	Hits	Earned Runs	Strike Outs	Home Runs	Wins	Losses	Outs Pitched	Batters Faced by Pitcher	Games Finished	Weight
0	1990	AbbottJim	23	185000	4.51	246	106	105	16	10	14	635	925	0	200
1	1990	AbbottPaul	23	100000	5.97	37	23	25	0	0	5	104	162	0	185
2	1990	AldredScott	22	100000	3.77	13	6	7	0	1	2	43	63	0	195
3	1990	AndersonAllan	26	300000	4.53	214	95	82	20	7	18	566	797	0	178
4	1990	AppierKevin	23	100000	2.76	179	57	127	13	12	8	557	784	1	180

Create Salary Brackets

```
In [3]: # Look at distribution of salaries (suppressing scientific notation)
df['Salary'].describe().apply(lambda x: format(x, 'f'))
```

Out[3]:

```
count      4937.000000
mean      3011304.443387
std       4265619.190449
min        100000.000000
25%       327000.000000
50%       980000.000000
75%      4000000.000000
max      33000000.000000
Name: Salary, dtype: object
```

```
In [5]: # create salary brackets and labels
bins = [0, 499999, 4999999, 9999999, 34999999]
labels = ['low', 'mid', 'high', 'top']
```

```
In [6]: # apply salary brackets
df['Salary Bin'] = pd.cut(df['Salary'], bins=bins, labels=labels)
df
```

Out[6]:

	Year	Full Name	Age	Salary	ERA	Hits	Earned Runs	Strike Outs	Home Runs	Wins	Losses	Outs Pitched	Batters Faced by Pitcher	Games Finished
0	1990	AbbottJim	23	185000	4.51	246	106	105	16	10	14	635	925	0
1	1990	AbbottPaul	23	100000	5.97	37	23	25	0	0	5	104	162	0
2	1990	AldredScott	22	100000	3.77	13	6	7	0	1	2	43	63	0
3	1990	AndersonAllan	26	300000	4.53	214	95	82	20	7	18	566	797	0
4	1990	AppierKevin	23	100000	2.76	179	57	127	13	12	8	557	784	1

	Year	Full Name	Age	Salary	ERA	Hits	Earned Runs	Strike Outs	Home Runs	Wins	Losses	Outs Pitched	Batters Faced by Pitcher	Games Finished
...
4932	2016	WorleyVance	29	2600000	3.53	84	34	56	11	2	2	260	365	13
4933	2016	WrightMike	26	510500	5.79	81	48	50	12	3	4	224	328	5
4934	2016	WrightSteven	32	514500	3.33	138	58	127	12	13	6	470	656	0
4935	2016	YoungChris	37	4250000	6.19	104	61	94	28	3	9	266	406	7
4936	2016	ZimmermannJordan	30	18000000	4.87	118	57	66	14	9	7	316	450	1

4937 rows × 15 columns

◀		▶
---	--	---

Encode Salary Bins column

```
In [7]: # encode object features
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
encoded_df = df.copy()
df['Salary Bin'] = le.fit_transform(df['Salary'])

df.head()
```

```
Out[7]:
```

	Year	Full Name	Age	Salary	ERA	Hits	Earned Runs	Strike Outs	Home Runs	Wins	Losses	Outs Pitched	Batters Faced by Pitcher	Games Finished	Weight
0	1990	AbbottJim	23	185000	4.51	246	106	105	16	10	14	635	925	0	200
1	1990	AbbottPaul	23	100000	5.97	37	23	25	0	0	5	104	162	0	185
2	1990	AldredScott	22	100000	3.77	13	6	7	0	1	2	43	63	0	195
3	1990	AndersonAllan	26	300000	4.53	214	95	82	20	7	18	566	797	0	178
4	1990	AppierKevin	23	100000	2.76	179	57	127	13	12	8	557	784	1	180

◀		▶
---	--	---

```
In [11]: # create new df using only top features
df = df.filter(['ERA', 'Batters Faced by Pitcher', 'Outs Pitched', 'Games Finished', 'Strike Outs', 'Salary Bin'])
df.head()
```

```
Out[11]:
```

	ERA	Batters Faced by Pitcher	Outs Pitched	Games Finished	Strike Outs	Salary Bin
0	4.51	925	635	0	105	1
1	5.97	162	104	0	25	1
2	3.77	63	43	0	7	1
3	4.53	797	566	0	82	1
4	2.76	784	557	1	127	1

Split Features/Target & Training/Testing Sets

Split into features and target

- **y variable:** Our target variable, Salary Bin
- **X variable:** Our features

In [12]:

```
# Split our preprocessed data into our features and target arrays
y = df["Salary Bin"].values
X = df.drop(["Salary Bin"],1).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=1)
```

C:\Users\alyss\anaconda3\envs\mlenv\lib\site-packages\ipykernel_launcher.py:3: FutureWarning: In a future version of pandas all arguments of DataFrame.drop except for the argument 'labels' will be keyword-only
This is separate from the ipykernel package so we can avoid doing imports until

Build and Instantiate StandardScaler object, then standardize numerical features

In [13]:

```
# Create a StandardScaler instance
scaler = StandardScaler()

# Fit the StandardScaler
X_scaler = scaler.fit(X_train)

# Scale the data
X_train_scaled = X_scaler.transform(X_train)
X_test_scaled = X_scaler.transform(X_test)
```

Build Neural Net Framework

HL1:

- 50 neurons
- activation fxn: relu

HL2:

- 40 neurons
- activation fxn: relu

HL3:

- 30 neurons
- activation fxn: relu

output layer:

- 4 neurons
 - same as number of salary bins, suggested from (<https://machinelearningmastery.com/loss-and-loss-functions-for-training-deep-learning-neural-networks/>)
- activation fxn: softmax
 - suggested for multiclass classification problems per (<https://machinelearningmastery.com/loss-and-loss-functions-for-training-deep-learning-neural-networks/>)

In [14]:

```
# Define the model - deep neural net
number_input_features = len(X_train[0])
hidden_nodes_layer1 = 50
hidden_nodes_layer2 = 40
hidden_nodes_layer3 = 4
```

```

nn = tf.keras.models.Sequential()

# First hidden layer
nn.add(
    tf.keras.layers.Dense(units=hidden_nodes_layer1, input_dim=number_input_features, activation="relu")
)

# Second hidden layer
nn.add(tf.keras.layers.Dense(units=hidden_nodes_layer2, activation="relu"))

# Third hidden layer
nn.add(tf.keras.layers.Dense(units=hidden_nodes_layer3, activation="relu"))

# Output layer
nn.add(tf.keras.layers.Dense(units=1, activation="softmax"))

# Check the structure of the model
nn.summary()

```

Model: "sequential"

Layer (type)	Output Shape	Param #
=====		
dense (Dense)	(None, 50)	300
dense_1 (Dense)	(None, 40)	2040
dense_2 (Dense)	(None, 4)	164
dense_3 (Dense)	(None, 1)	5
=====		
Total params: 2,509		
Trainable params: 2,509		
Non-trainable params: 0		

Compile the Model

```

In [15]: # Compile the model
nn.compile(loss="CategoricalCrossentropy", optimizer="adam", metrics=["accuracy"])

```

Train the model

```

In [16]: # Train the model
fit_model = nn.fit(X_train,y_train,epochs=200)

```

Epoch 1/200
116/116 [=====] - 0s 643us/step - loss: 0.0000e+00 - accuracy: 0.3787
Epoch 2/200
116/116 [=====] - 0s 600us/step - loss: 0.0000e+00 - accuracy: 0.3787
Epoch 3/200
116/116 [=====] - 0s 617us/step - loss: 0.0000e+00 - accuracy: 0.3787
Epoch 4/200
116/116 [=====] - 0s 617us/step - loss: 0.0000e+00 - accuracy: 0.3787
Epoch 5/200
116/116 [=====] - 0s 591us/step - loss: 0.0000e+00 - accuracy: 0.3787
Epoch 6/200
116/116 [=====] - 0s 609us/step - loss: 0.0000e+00 - accuracy: 0.3787
Epoch 7/200
116/116 [=====] - 0s 617us/step - loss: 0.0000e+00 - accuracy: 0.3787
Epoch 8/200
116/116 [=====] - 0s 609us/step - loss: 0.0000e+00 - accuracy: 0.3787
Epoch 9/200
116/116 [=====] - 0s 609us/step - loss: 0.0000e+00 - accuracy: 0.3787

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

Epoch 200/200
116/116 [=====] - 0s 618us/step - loss: 0.0000e+00 - accuracy: 0.3787

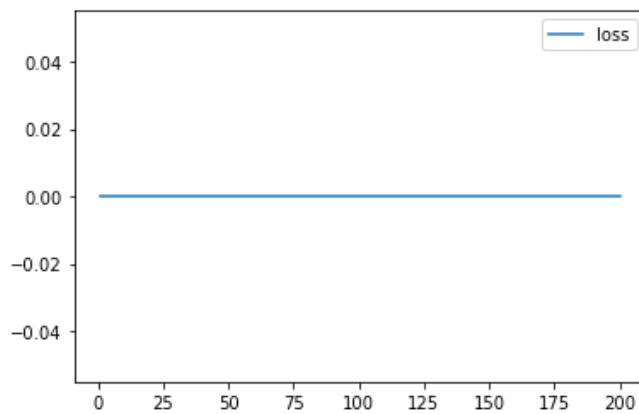
```
In [17]: # Evaluate the model using the test data
model_loss, model_accuracy = nn.evaluate(X_test_scaled,y_test,verbose=2)
print(f"Loss: {model_loss*100:.2f}%, Accuracy: {model_accuracy*100:.2f}%")
```

39/39 - 0s - loss: 0.0000e+00 - accuracy: 0.3887 - 121ms/epoch - 3ms/step
Loss: 0.00%, Accuracy: 38.87%

```
In [18]: # Create a DataFrame containing training history
history_df = pd.DataFrame(fit_model.history, index=range(1,len(fit_model.history["loss"])+1))

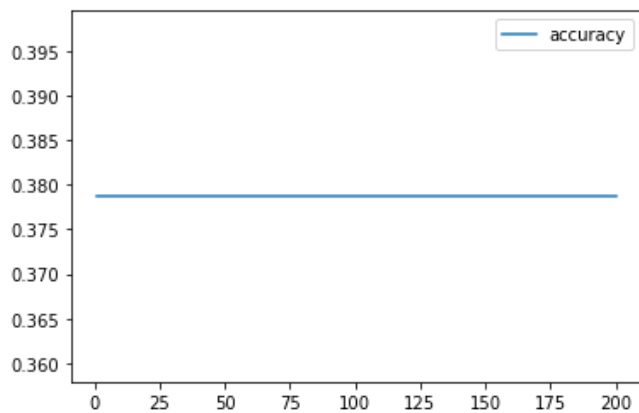
# Plot the loss
history_df.plot(y="loss")
```

Out[18]: <AxesSubplot:>



```
In [19]: # Plot the accuracy
history_df.plot(y="accuracy")
```

Out[19]: <AxesSubplot:>



In []: