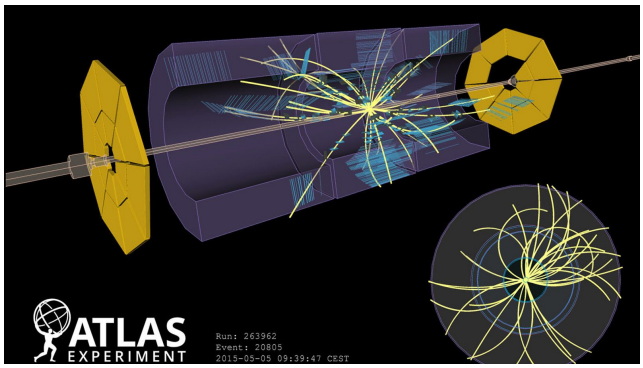
Predicting Characteristics of Dielectron Events in CERN High Energy Collisions



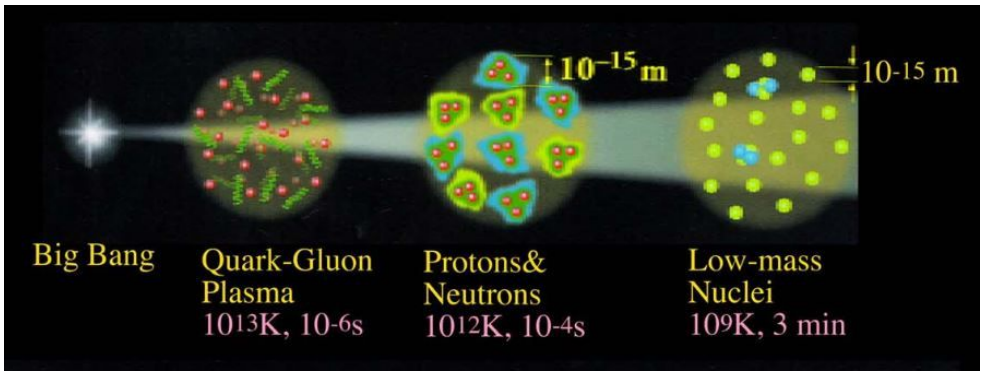
Tony Lin and Jesse Cox

https://www.youtube.com/watch?v=bTHzB4h0po4

Large Hadron Collider

- The Large Hadron Collider (LHC) is a particle accelerator located around Geneva, Switzerland, that smashes together the nuclei of two atoms at near light speed.

Quark-Gluon Plasma



- The resulting reaction is hot enough for protons and neutrons melt, such that their constituent "particles," flow freely in a unique form of matter called "Quark-Gluon Plasma" (QGP), which resembles the state of the universe shortly (nanoseconds) after the big bang.

- In QGP, particles called hadrons are created and survive for less than a trillionth of a second before decaying and *occasionally* producing two electrons, in what is called a "dielectron event."

- Studying dielectron events gives physicists some information about the energy profile of decaying hadrons, which in turn give us clues about the very early stages of the universe.

Data

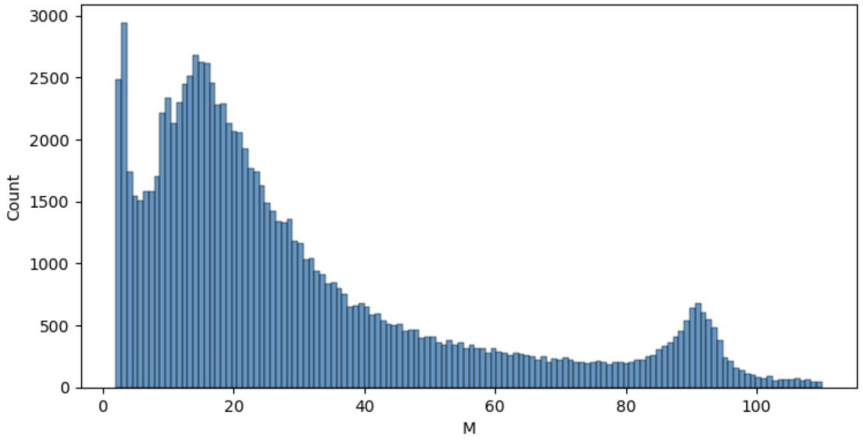
	Run	Event	E1	px1	py1	pz1	pt1	eta1	phi1	Q1	E2	px2	py2	pz2
0	147115	366639895	58.71410	-7.311320	10.531000	-57.29740	12.82020	-2.202670	2.177660	1	11.28360	-1.032340	-1.88066	-11.077800
1	147115	366704169	6.61188	-4.152130	-0.579855	-5.11278	4.19242	-1.028420	-3.002840	-1	17.14920	-11.713500	5.04474	11.464700
2	147115	367112316	25.54190	-11.480900	2.041680	22.72460	11.66100	1.420480	2.965600	1	15.82030	-1.472800	2.25895	-15.588800
3	147115	366952149	65.39590	7.512140	11.887100	63.86620	14.06190	2.218380	1.007210	1	25.12730	4.087860	2.59641	24.656300
4	147115	366523212	61.45040	2.952840	-14.622700	-59.61210	14.91790	-2.093750	-1.371540	-1	13.88710	-0.277757	-2.42560	-13.670800
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
99995	146511	522575834	12.31310	-10.658000	5.164440	3.36858	11.84330	0.280727	2.690370	-1	1.80181	0.668609	-1.58437	0.537805
99996	146511	522786431	18.46420	7.854990	15.133000	-7.08659	17.05020	-0.404510	1.092010	1	14.69110	-1.418020	-2.28117	-14.443500
99997	146511	522906124	4.18566	-3.273500	-0.308507	-2.59013	3.28801	-0.723075	-3.047630	1	72.81740	-11.074900	-9.28179	-71.369300
99998	146511	523243830	54.46220	11.352600	11.880900	51.92400	16.43280	1.867800	0.808132	-1	8.58671	0.378009	3.07828	8.007050
99999	146511	524172389	7.64000	0.886162	5.478900	-5.25033	5.55010	-0.842662	1.410440	1	52.10880	16.807500	-4.60510	49.108400

100000 rows × 19 columns

https://opendata.cern.ch/record/304

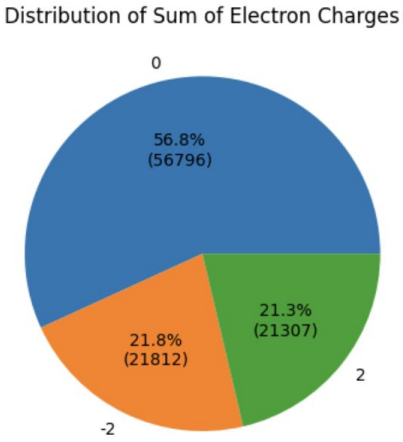
- Our dataset is 100 dielectron events from 13 high energy “runs” of the LHC.
- We have features that represent the physical state of both electrons associated with the event. These include:
    - Momentum in three directions (mention that these are the p_ features).
    - Mass.
    - Electrical Charge.
    - Angle of Momentum from Beam Axis.

Distributions



- All features demonstrated some non-normal distribution, and often had a complex modality, so there seemed to be some possibility of predicting some classification.

Charge Combinations & Baseline Model



- When performing exploratory data analysis, our curiosity piqued at the fact that electrons produced in hadron decay aren't always negatively charged. Just over half of our dataset contains events that produced both a negatively and positively charged electron (known also as a positron). Then, among the events with similarly-signed electrons, seeing two negatively or two positively charged electrons seemed equally likely.

- We decided to run experiments that attempt to classify the sum of the charges of the electrons in each event as "zero-sum," a "double-negative," or a "double-positive" event.

- Our choice for a baseline classifier is to select the majority class. That is, we simply predict that the two electrons produced in hadron decay will have opposing charges.

    - The result is an accuracy of roughly 56.8%.

Random Forest

'n_estimators': [100, 200, 300]

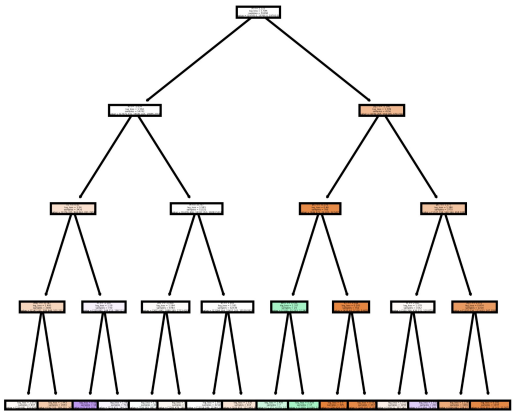
'max_depth': [None, 10, 20, 30]

'min_samples_split': [2, 5, 10]

'min_samples_leaf': [1, 2, 4]

'max_features': ['auto', 'sqrt']

'bootstrap': [True, False]



KNN

'n_neighbors' : [10,20,30,50,100,200,300,400,500]

CNN

'learning_rate': [0.1, 0.01, 0.001]

'epoch':[10, 20, 30, 50, 100, 200]

'batch_size':[26, 32, 64, 128]

Layer (type)	Output Shape	Param #
dense_20 (Dense)	(None, 128)	2,304
dropout_8 (Dropout)	(None, 128)	0
dense_21 (Dense)	(None, 64)	8,256
dropout_9 (Dropout)	(None, 64)	0
dense_22 (Dense)	(None, 32)	2,080
dense_23 (Dense)	(None, 3)	99

Experiments

Baseline (Majority Class): Accuracy = 56.68%

Best Random Forest: {'n_estimators': 100, 'min_samples_split': 5, 'min_samples_leaf': 1, 'max_features': 'sqrt', 'max_depth': None, 'bootstrap': True}

Accuracy = 55.15%

Best KNN: {'n_neighbors': 200, 300, 400}

Accuracy = 56.68%

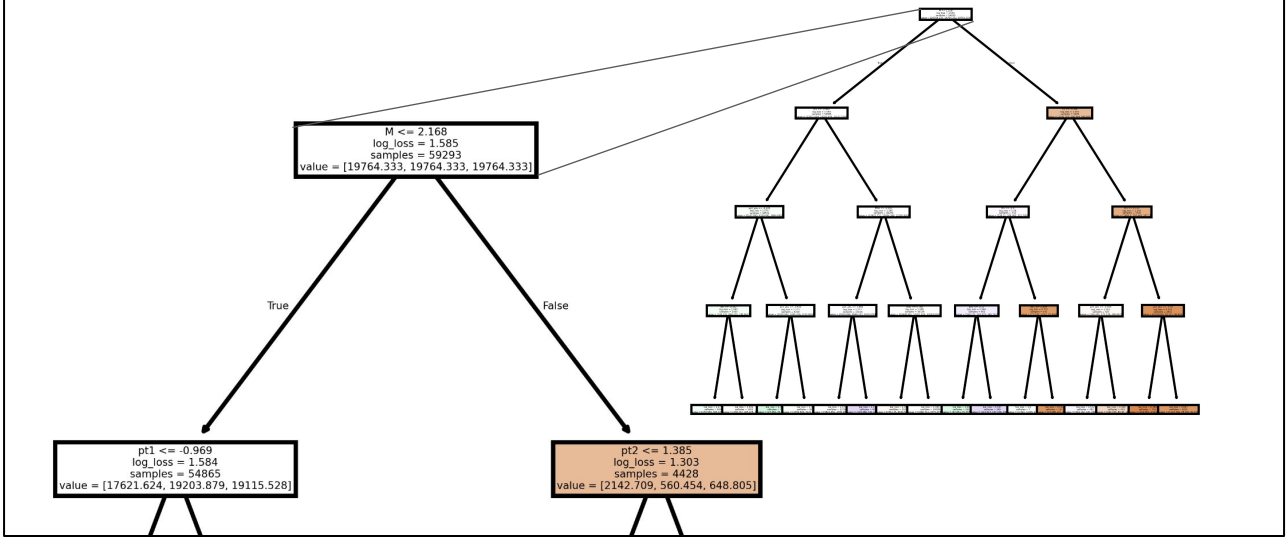
Best CNN: {'learning_rate':0.001, 'epoch':20, 'batch_size':20}

Accuracy = 56.68

Conclusions

- Interestingly, all of the models' accuracies maxed out at 56.86% (Exact same as the baseline)
- Even though there is visible pattern in the EDA, none of the models seem to have a strong prediction
- We may need much more prior knowledge in this field to perform feature engineering on this dataset

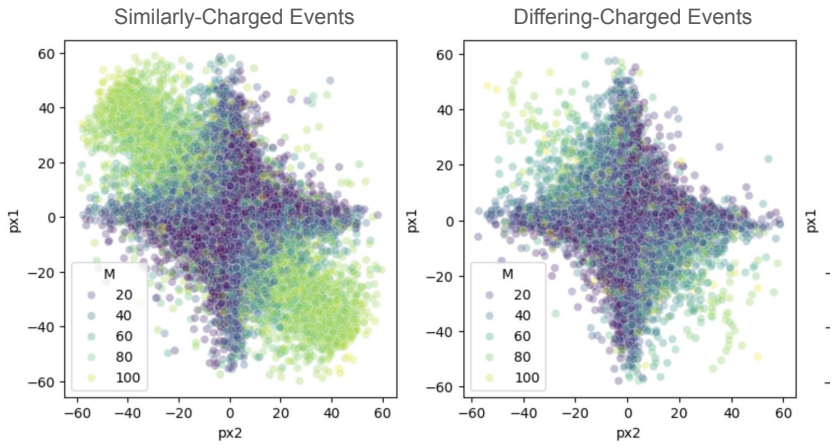
Conclusions



- While the accuracy of predicting the class of charge distribution may seem to be limited to the accuracy of a majority-class prediction, there was still *some* interesting information gleaned from our Random Forest model.

- On creating a graph of the result of our random forest algorithm (seen on this slide), we noticed that the first decision was always made on the invariant mass of the system. This creates trees containing what we may call “low-mass events” and “high-mass events,” where the high-mass tree seems to carry more decision-pertinent information than the low-mass tree. [Mention coloring of the tree in slide]

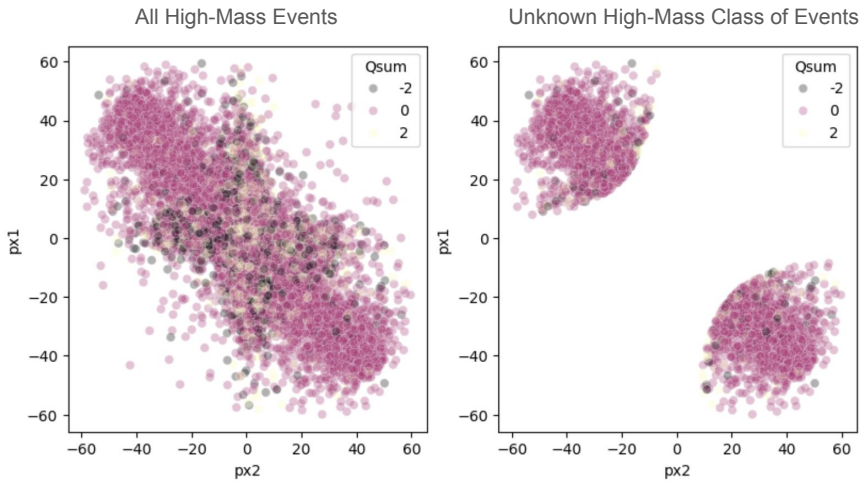
Conclusions



- Digging into this further, we segregate the dataset into same-charge and differing-charge events, and then color by Invariant Mass while plotting against some aspect of the electrons' physical state (here we have px1 and px2, which the momentum of either electron in the x direction).

- We then begin to see the emergence of some class of high-mass event that is far more consistently similarly-charged than low mass events.

Conclusions



- Here we can see that class of consistently zero-sum events occur when the product of px1 and px2 is a large negative number.
- However, if we isolate that new class, and always predicting it as a “zero-sum” event (same charge), and then go on to do another majority class prediction for the remaining data, this process still gives us the same overall accuracy, because the remainder of the dataset is more or less equally distributed between classes.
    - In other words, all the information gain that results in the 56% accuracy (rather than 50%) in the majority class prediction is a result of this new class of event.
- This tracks to the low-information / high-information tree-structure in our random forest model.

Contributions