CryptoDataScience Introduction Post

Kaggle Competitions, Ensemble Methods and examining the relevancy of a century old dataset to Cryptocurrency

Kaggle

- The Main Community of Data Scientists
- Reward based competitions
- Learning based competitions

- We will be participating in the most popular competition
- The dataset is great for introduction EDA, and building your first model whether it be logistic regression, or a decision tree.
- We will use Ensemble Methods to succeed and not succumb to some typical pitfalls people fall into trying to achieve a higher score

What does the Titanic have to do with Cryptocurrency?

- Before we answer that question its important as Data Scientists to evaluate:

- Broader Context
  - What is the Historical Relevance of the Titanic?
  - Extracting all value from your businesses data is something Senior Data Scientists should do
- Root Cause of the Problem
  - Why were there maritime casualties?

Historical Relevance of Titanic

- Largest and Most Luxurious Ship of Its Time making it incredibly popular amongst the most wealthy in Britain
- Had Several New Technologies
  - Separate Watertight compartments that would enable the boat to float in an event of a water breach
  - Wireless technology that allowed Morse code messaging over long distances
- Part of a broader migration movement to the New World
- After the fact, its sinking accelerated naval safety regulation

Why did the Titanic Sink?
(Root Cause Investigation)

Investigations in Britain and America found:

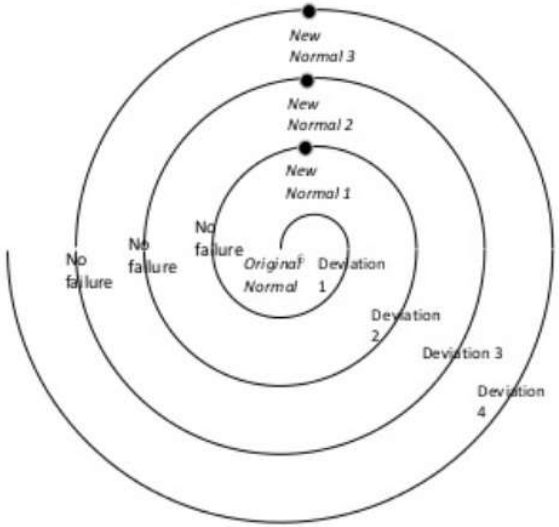
- Insufficient number of Lifeboats
- Speed was Excessive
- Complacency, ship was thought to be unsinkable

Why did the Titanic Sink?
(Root Cause Investigation)

- Normalization of Deviance
- Term created by Sociologist Diane Vaughan, in the 1980s to characterize what happened to Shuttle Challenger
- Engineers warned O-rings could cause catastrophic malfunction
- Management culture around safety was relaxed due to urgency to take flight

Normalization of Deviance

The "Deviation Spiral"

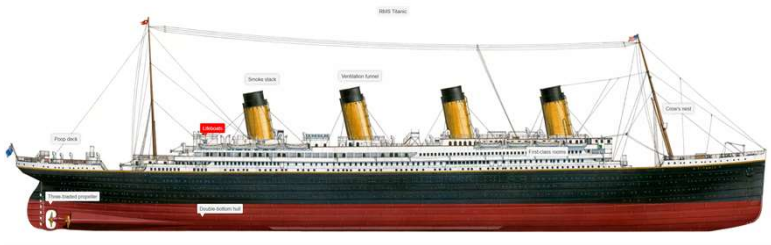


How is the Titanic related to Cryptocurrency?

CryptoCurrency is Currency

=

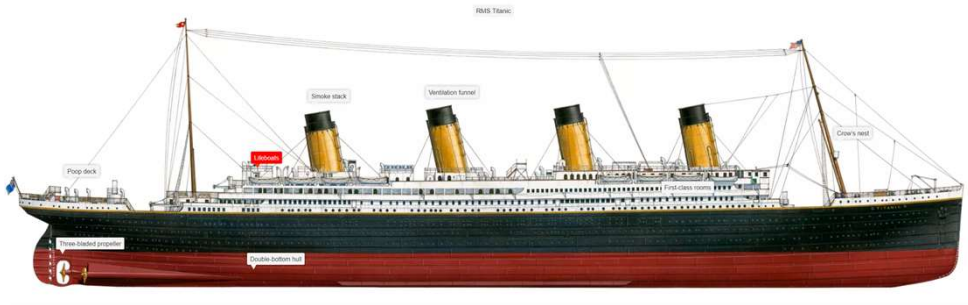
How is the Titanic related to Cryptocurrency?



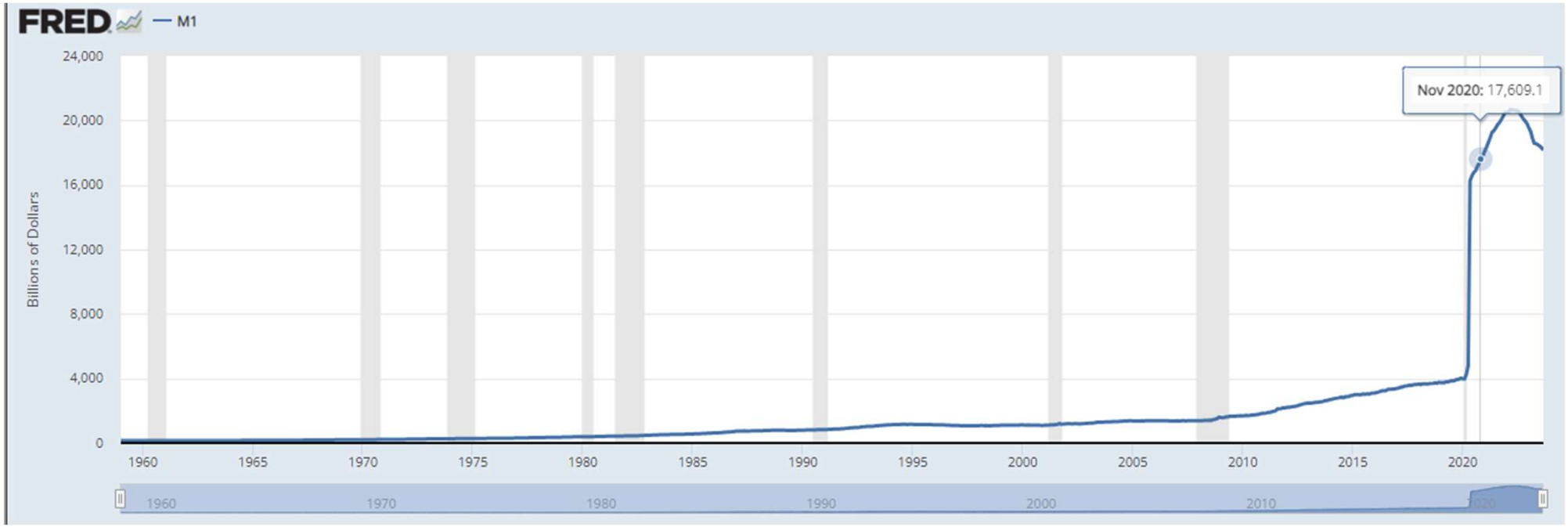
Cryptocurrency is the futuristic form of currency.

Let’s look at the current version of currency:

Currency Today

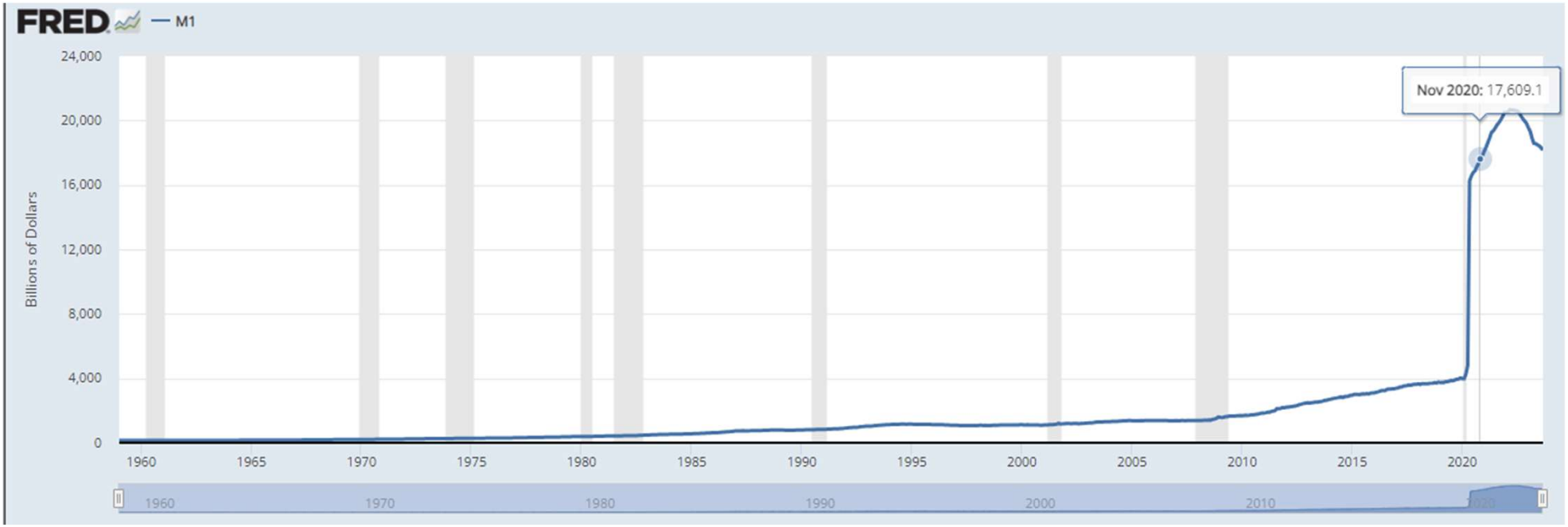


Is this a Glacier or Currency Today?



https://fred.stlouisfed.org/series/M1SL

Just who or what is Fred M1? And why is there such a sharp increase?



SA

FRED: Federal Reserve Economic Data

What is M1?
Why is there such a sharp increase?

Let’s ask pplx.ai!

What is Federal Reserve M1, what is the definition, why was there a sharp increase in Spring 2020

Sources

What's Driving Up Money Growth? - Liberty Street... newyorkfed • 1

[PDF] Inflation in the U.S. Economy: Causes and... congress • 2

The Fed's Policies since the 2020 Coronavirus... mises • 3

View 2 more

Answer

The Federal Reserve M1, or "money stock," includes currency in circulation, demand deposits, and other liquid assets. The sharp increase in M1 in the spring of 2020 was primarily due to the economic impact of the COVID-19 pandemic and the response of the Federal Reserve. The Fed's asset purchases and other policy measures led to a significant increase in reserves, which in turn contributed to the growth in M1. This expansion of the money supply was aimed at providing liquidity and support to the economy during a period of severe disruption [1] [3].

The Federal Reserve (Fed) defines stable prices to be inflation of 2% annually [2]. The increase in M1 was a result of the Fed's policies in response to the 2020 Coronavirus Panic, which led to a significant rise in reserves and money growth [3]. The specific details of the increase in M1 and its correlation with Fed policies can be found in the provided sources.

Share Rewrite

what is the impact of increase in m1 on economy

Sources

M1 Money Supply: How It Works and How to...
investopedia · 1

What's behind the recent surge in the M1 money...
stlouisfed · 2

Money Supply Definition: Types and How It Affect...
investopedia · 3

View 2 more

Answer

An increase in the M1 money supply can have several impacts on the economy. When the money supply increases, it typically lowers interest rates, which encourages more investment and stimulates consumer spending. This, in turn, can lead to increased business activity, higher demand for labor, and a rise in production. However, an excessive increase in the money supply can also lead to inflation as the demand for goods and services rises, causing prices to increase [3] [4].

The recent surge in the M1 money supply in response to the COVID-19 pandemic was aimed at making it easier to access capital, stimulating the economy, keeping workers employed, and encouraging business activity [1]. The increase in M1 was largely due to the Federal Reserve's policies and the response of banks to accommodate the heightened demand for money during this period [2]. Overall, while an increase in the M1 money supply can provide a boost to the economy, it's important to manage it carefully to avoid potential negative consequences such as inflation.

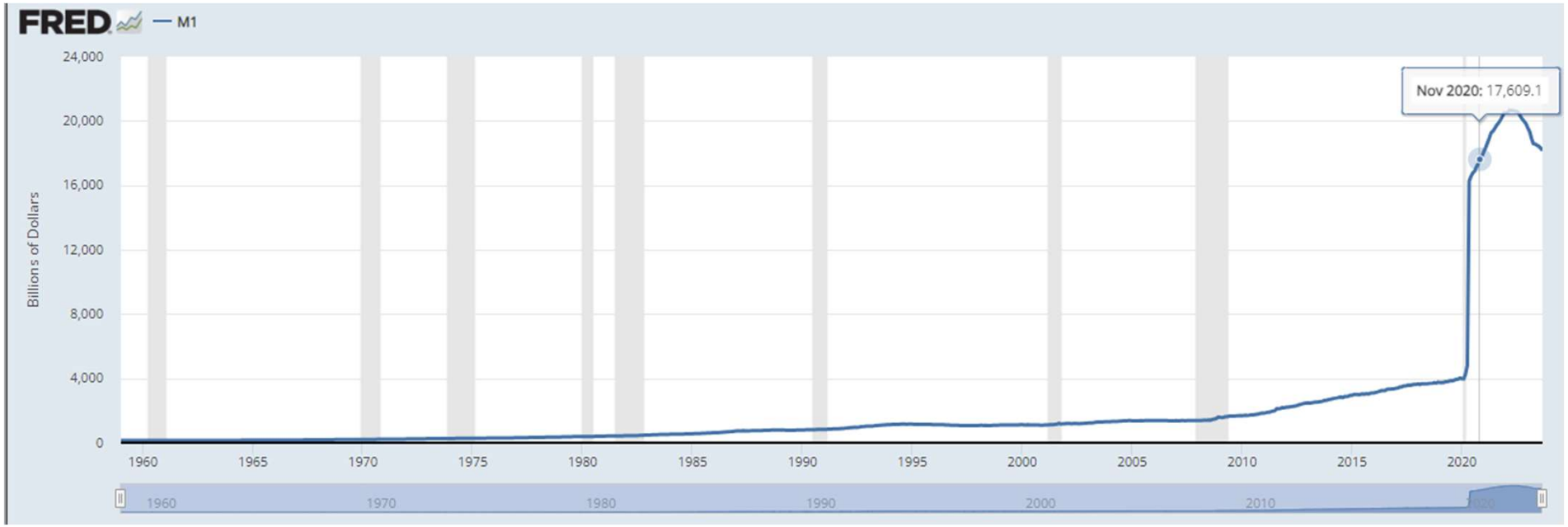
Share Rewrite

Statement from Federal Reserve:

- In late February and early March of 2020, the Fed cut its policy interest rate dramatically to help ease credit conditions during the COVID-19 crisis. The resulting acceleration in the supply of M1 can be understood largely as banks accommodating an increase in people's demand for money.

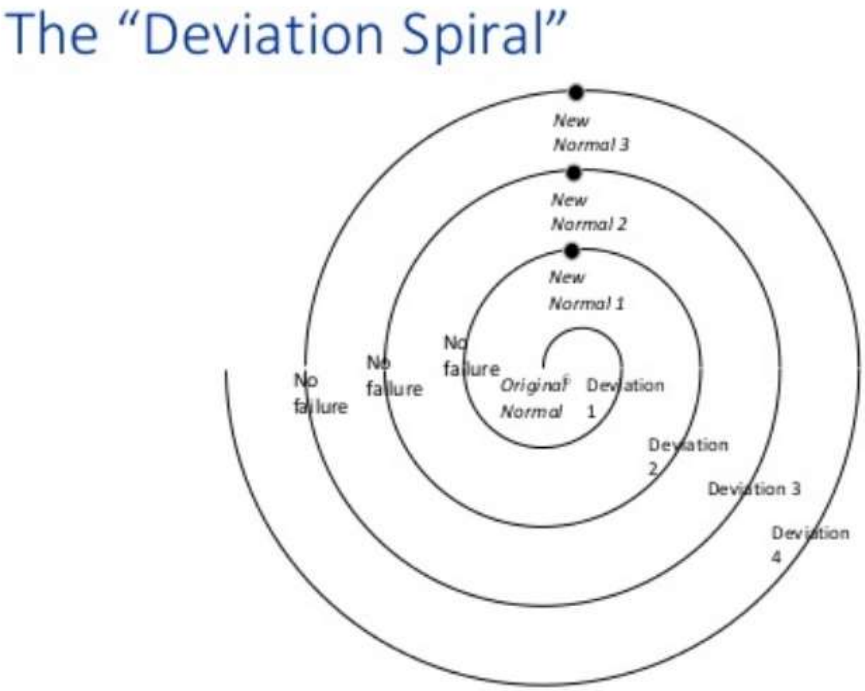
https://fredblog.stlouisfed.org/2021/01/whats-behind-the-recent-surge-in-the-m1-money-supply/

For people around the world,
Is this a Glacier?



Inflation the Glacier

Is there a possibility this is relevant to our current Financial system?



What some in the industry are saying:

- Bitcoin was created for a moment like this. Inscribed into its Genesis Block is the phrase “Chancellor on brink of second bailout for banks,” an homage to the government bailouts of 2008 and the last great financial crisis.

- The call out is a subtle nod to the need for a sovereign form of money without any central intermediary.

- As the US government turns to slashing interest rates, passing large stimulus packages, and infinite quantitative easing, Bitcoin will soon do the opposite in the next Bitcoin halving. The contrast could not be more stark.

How to Buy Bitcoin

- https://www.coinbase.com/how-to-buy/bitcoin

- https://www.kraken.com/learn/buy-bitcoin-btc

Machine Learning Building Steps

- 1) Problem Identification
- 2) Data Collection
- 3) Data Preparation
- 4) Model Selection
- 5) Model Training/ 6)Hyperparameter Tuning
- 7) Model Evaluation
- Repeating Steps 5-7

Problem Identification

Our First step, which is already done for us is to understand or define the problem

- In this case, the problem has been defined back in 2012, as the dataset was used in a Kaggle competition.
- Problem Definition: Predicting Life or Death in the RMS Titanic
- The Titanic was a luxury British steamship that sank on its maiden voyage after hitting an iceberg on April 1912. The ship had lifeboats, but not enough for everyone.
- Among the 2240 total passengers 1517 perished. And although there was some element of luck in getting on a lifeboat, some groups were more likely to get on a lifeboat than others.
- In this notebook we will try to create a model that determines whether or not someone was likely to survive using passenger data (ie name, age, gender, socio-economic class, etc) collected prior to embarking.

Data Collection

- Kaggle Provided Data
- 1309 Total Records, 891 records in train.csv, 418 records in test.csv
- For the sake of the competition we won’t train on any test data
- A lot of Kaggle notebooks make this mistake, and get a higher score because of it, (https://www.kaggle.com/competitions/digit-recognizer/discussion/296110)

List of Variables

- PassengerId
- Survived
- Pclass
- Name
- Sex
- Age
- SibSp
- Parch
- Ticket
- Fare
- Cabin

Data Preparation

- With the exception of some missing variables the Dataset is Tidy
- A row represents what it should, data is one place
- Data Preparation consists of Feature Engineering, Variable by Variable
- Creating dummy variables
- Scaling Variables

Feature Engineering

Variable by Variable

Explanatory Variable #1

- Name

- Extracted the Title (Mr., Mrs. etc)
- Titles that weren't conventional all had a connotation of status, by Occupation, or by Birth. (Doctor, Colonel, Countess etc.)

Explanatory Variable #2

- Sex

- This variable did not need any transformation

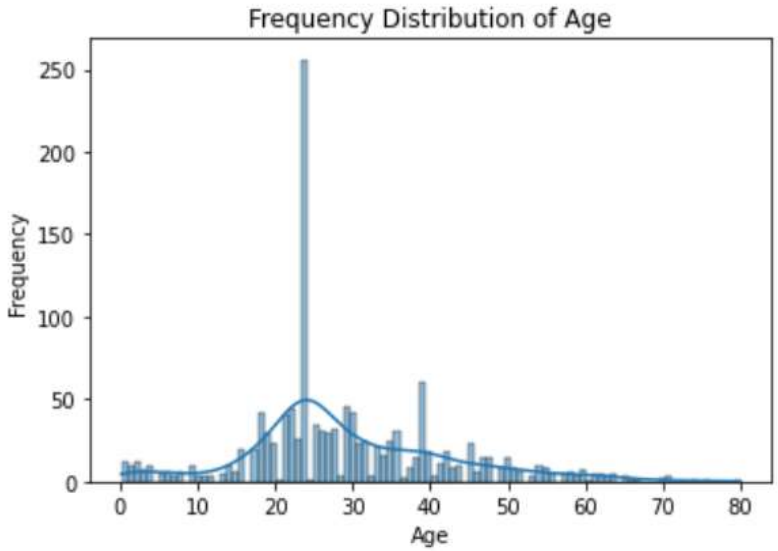
Explanatory Variable #3

- Pclass – Passenger Class

- This variable did not need any transformation

Explanatory Variable #4

- Age



- Imputed missing Age values by calculating median Age by Class.
- Created Categorical Variable by what Age Group they were in.(Children Teens, Young Adults etc)

Explanatory Variable #5 and #6

- SibSp, Parch

- Combined these two numeric values to calculate Total Number of family members

Explanatory Variable #7

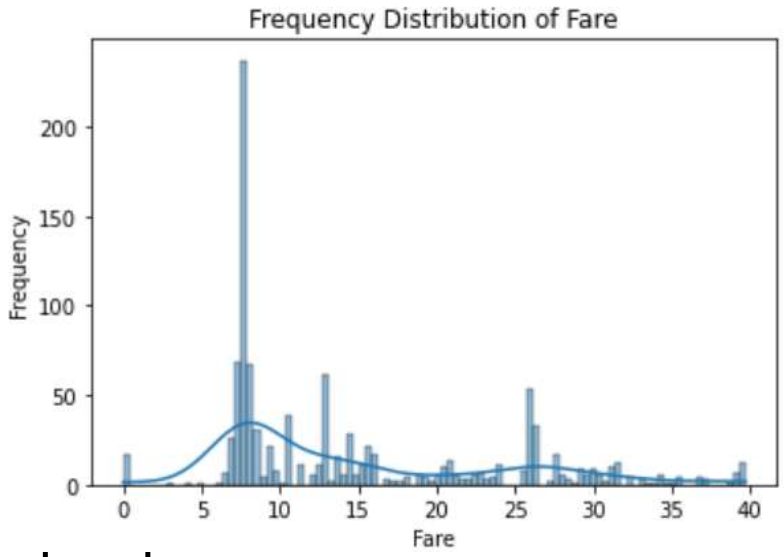
- Ticket

- Reduced the amount of groups by putting less common tickets into higher level groups
- Used to create a new feature that was effectively the Number of other passengers with same Ticket

Credit goes to https://www.kaggle.com/maciejsabalski for this code.

Explanatory Variable #8

- Fare



- Bucketed the Fare based on different price levels

Explanatory Variable #9

- Cabin

- Created a Yes/no field based on whether or not they had a cabin

Explanatory Variable #10

- Embarked

- Missing Embarked was given most popular Embarking point, Southampton.

Machine Learning Building Step 4 -7

- Model Selection

Why Ensemble Methods in General

- Significant number of Kaggle competitions won by Ensemble methods
- Ensemble Methods combine multiple models to produce more accuracy
- Ensembles, average out errors, and can reduce risk of overfitting
- Different models may perform better on different subsets
- More robust to noise and outliers
- Scale computing power very well

List of Models

- Tree based Boosting models – ADABoost, XGBoost
- Tree based Bagging model – Random Forest
- SVM – Hyperplane model

Model Training Tuning Evaluation

- Model Training, Tuning is done via GridSearchCV
- Evaluation is done by submitting the dataset to the website

Why XGBoost

- Decision Tree based Ensemble
- Introduced more recently
- XGBoost builds Complex Trees, not just Stumps
- XGBoost also focuses on training instances that were misclassified.
- **XGBoost** uses gradient descent on a specified loss function, which allows for more nuanced adjustments
- Has more parameters including learning rate, regularization

XGBoost Result

- Very Good Performance on the Training Set – 90%
- Not Great Performance on the Test Set – 76.3%
- Due to its use of complex trees this model does even better on the Training Set, yet worse on the Test set due to overfitting

Why ADABoost

- Decision Tree based Ensemble
- ADABoost builds weak learners in sequence (Stumps)
- Learners focus on examples that were hardest to predict by previous learner
- Misclassified training samples are more likely to be picked up by the new learner in the next round of training

ADABoost Result

- Very Good Performance on the Training Set – 83%
- A Decent performance on the Test Set but not as high as Training– 77%
- Overfitting

Why Random Forest?

- Decision Tree based Ensemble
- Bagging, or Bootstrap Aggregation, as opposed to Boosting
- Trees are built in parallel using different samples of data.
- Trees are built to full depth
- Trees are averaged to reduce variance

Random Forest Result

- Very Good Performance on the Training Set – 84.6%
- Not as great Performance on the Test Set – 77.5%
- Due to its use of complex trees this model does even better on the Training Set, yet worse on the Test set due to overfitting

WHY SVM?

- Not an Ensemble Method
- Very different than a tree based model – Uses a hyperplane that represents the largest separation between the two classes.
- Good complement to Tree based model in our final Voting Classifier
- Parameter tuning includes Kernel

SVM Result

- Very Good Performance on the Training Set – 84.2%
- The best Performance on the Test Set so far, but still overfit – 77.7%

Why Voting Classifier?

- Easy way to combine different individual models, or ensemble models
- We combine our to best models Random Forest and SVM for the Final Model
- RF and SVM are correlated

SVM and Random Forest

- We combine our SVM and Random Forest in a Hard Voting Classifier Model
- Hard Voting Classifier means the Label with the most votes from the “Voting Models” is the Final Predicted Label
- Soft Voting Classifier means the “Voting Models” predicted probabilities are averaged to come up with the Final Predicted Label
- 79.9% accuracy on Test Set

What Titanic scores mean 70-100%

- 70-77% - A lot of beginner models end up with this score in the Kaggle competition, it is a very respectable score
- 77-80% - Good Feature Engineering, Fine tuning a model, picking a good general model
- 82% + Models that fall in this category, didn't split up the Test, Training data in the same way as other contestants (https://www.kaggle.com/competitions/digit-recognizer/discussion/296110)
- Also models of 81% + have …Data Leakage, for example Giving Survival Leakage by attaching the survival results of the rest of the family

Final Notes about the ML Exercise

- Historical Dataset Limitations
- Will never be able to apply this model to new data

- Data Science requires a lot more Data Engineering than what was involved, a lot more intuition and experience that comes from working with different datasets

Happy 15th Birthday to Cryptocurrency!

The Genesis Block was Mined January 3, 2009!