Scalable Knowledge Sharing Platform; An Artificial Intelligence Approach

by: Ibrahim O. Odufowora

Outlines

- Problem Description
- Understanding the data set
- Data Exploration
- Parameter Tuning
- Base Model
- Activities Update

Problem Description

- Identification of duplicate questions.
- Classification Problem
- Supervised Learning
- 2 important features - (Text in nature)
- 1 response variable - (Binarized)
- P << n, then most models can be used.
- No missing data

Data Exploration - Word Cloud

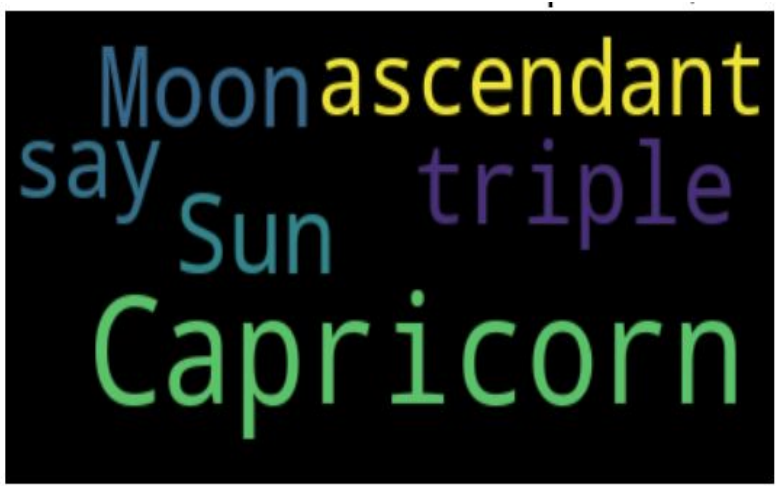
What is the step by **step** **guide** to **invest** in **share** **market** in **india**?

What is the **step** by step **guide** to **invest** in **share** **market**?

Data Exploration - Word Cloud

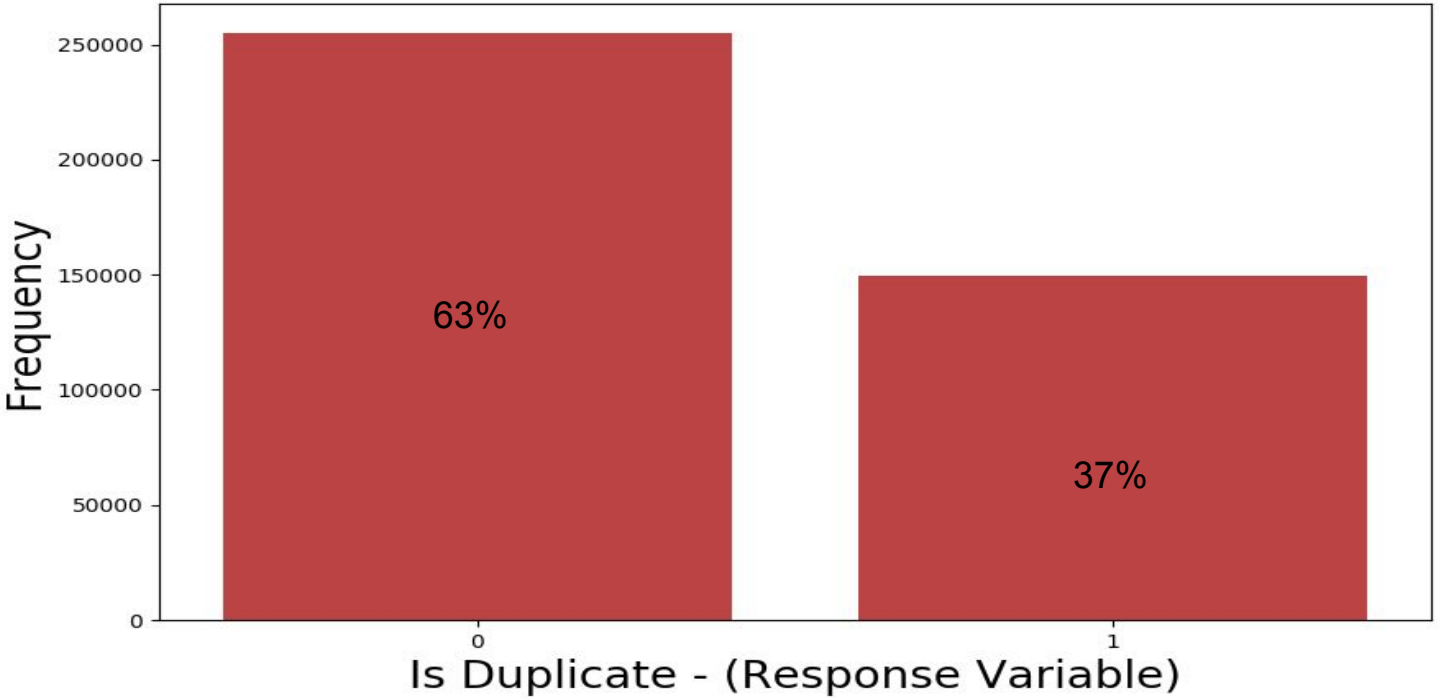
Astrology: I am a **Capricorn** **Sun** Cap **moon** and **cap** **rising**...what does that **say** about me?

I'm a **triple** **Capricorn** (**Sun**, **Moon** and **ascendant** in Capricorn) What does this **say** about me?

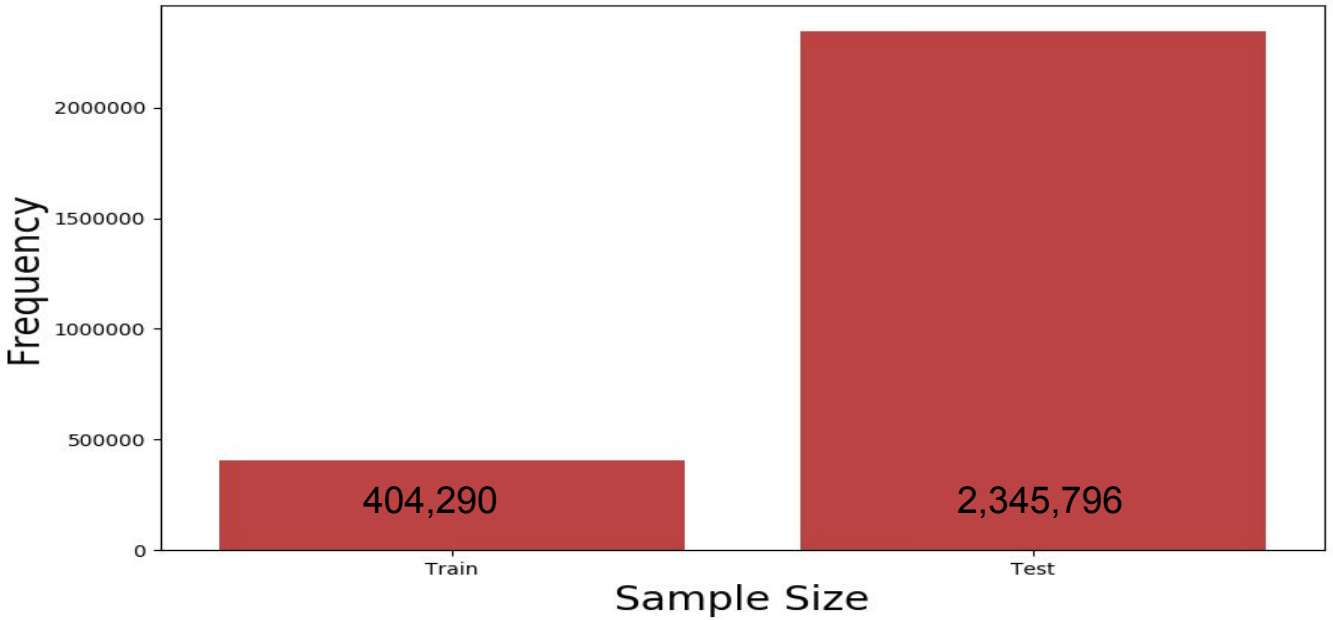


Data Exploration

Pr(y = 1 | random guess) = 0.37



Data Exploration



Parameter Tuning

- Cross-validation or GridSearch would also be used to select best model parameters.

- Given the imbalance in the frequency distribution of response predictor, it might be suggested to use stratified random sampling method during CV.

Base Model

Model	Train Accuracy
Naive Bayes	53.2%

Activities Update

- Completed Task
  - Data Analysis
  - Data Preprocessing
- On-Going Task
  - Training Models
  - Automating parameter tuning
  - Improve on the base model
- Not Started
  - Test Models
  - Integrate with API

References

- Quora. 2017. *First Quora Dataset Release: Question Pairs*. [ONLINE] Available at: https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs. [Accessed 30 September 2018].


Thank
You!

Outlines

- Problem Description
- Understanding the data set
- Data Exploration
- Solution Methodology & Implementation
  - Data preprocessing
  - Model building & Parameter tuning
  - Model Validation
- Environment Setup
  - Development
  - Production
- References

Problem Description

- Identification of duplicate questions.
- A page for each logically distinct question.
- Classification Problem
- Supervised Learning
- P << n, then most models can be used.
- A ‘scalable question platform’ should provide answers to each distinct question in a single resource location. This will enhance proper resources utilization.

Solution Methodology & Implementation

- Methodology:
  - Automate the process of detecting semantically equivalent questions.
  - Direct a user to an existing resource/page if the question already exist.
- Implementation:
  - Build effective and scalable machine learning and natural learning processing models.
  - Data preprocessing.
  - Models building & Parameters tuning.
  - Models Validation.
  - Interface a scalable, online real-time RESTful API with the best model.
  - Load balance the API in order to minimize response time and to avoid overload.

Environment Setup

- Development
  - Build the models on the train set.
  - Use test set for models validation.
  - Interface a Flask RESTful API with the best model.
- Production
  - Move the best model and its optimal parameters to production.
  - Interface a RESTful API with WSGI server with the model parameters.
  - Load balance the API.
  - As more data suffaces, rebuild the model occasionally to improve predictability.