Final Project Report of Internship Program

On

“Blood Donation Prediction”

MedTourEasy, New Delhi

28th September, 2023

Sai Karteek Kompally

Acknowledgement

I would like to express my heartfelt gratitude to MedTourEasy for granting me the incredible opportunity to serve as a Data Analyst Intern. It is with great enthusiasm and appreciation that I acknowledge the trust and confidence you have placed in me.

This internship experience has been nothing short of transformative, and I owe a substantial part of my growth to the unwavering support and guidance of Mr. Ankit Hasija, my mentor. His mentorship has been invaluable, providing me with insights, knowledge, and direction that I will carry with me throughout my career. Mr. Hasija's dedication to fostering my professional development has truly made a significant impact on my journey as a Data Analyst.

I am deeply thankful to the entire MedTourEasy team for the warm welcome and the enriching work environment. It has been a privilege to be a part of such a dynamic and innovative organization, where I have had the opportunity to apply my skills and knowledge to real-world challenges.

I look forward to continuing my journey with MedTourEasy, and I am excited to contribute further to the company's mission. Once again, thank you for this remarkable opportunity and for believing in my potential.

Abstract

The project titled "Blood Donation Prediction" aims to address a critical need in the healthcare domain by leveraging data-driven techniques to predict blood donation patterns. Blood shortages pose a significant challenge to healthcare systems worldwide, impacting patient care, emergency response, and medical procedures. This project harnesses the power of data analysis and machine learning to enhance blood donation management, ensuring a steady and sufficient blood supply.

A blood transfusion is a way of adding blood to the body after an injury or illness. Blood transfusion saves lives - from replacing lost blood during major surgery or a serious injury to treating various illnesses and blood disorders. Ensuring that there's enough blood in supply whenever needed is a serious challenge for the health professionals. The demand for blood fluctuates throughout the year. As one prominent example, blood donations slow down during busy holiday seasons. An accurate forecast for the future supply of blood allows for an appropriate action to be taken ahead of time and therefore saving more lives.

In conclusion, this project addresses a critical societal challenge by harnessing the power of data and technology. By predicting blood donation events with greater accuracy, this project has the potential to save lives and ensure that healthcare systems have a reliable supply of blood when needed the most. It represents a promising step towards a more efficient and data-driven approach to blood donation management.

Table of Contents

Contents

1. Introduction

1.1 About the Company

MedTourEasy is not just an online medical tourism marketplace; it's your gateway to a world of healthcare possibilities. We are a global healthcare company dedicated to providing you with the essential information to evaluate your healthcare options globally. Our mission is to assist you in finding the perfect healthcare solution tailored to your specific needs, all while ensuring affordability and maintaining the highest quality standards that you rightfully expect in healthcare.

At MedTourEasy, we believe that healthcare should know no boundaries. That's why we are committed to improving access to healthcare for people all over the world. Our platform is designed to be user-friendly, making it easy for patients to seek medical second opinions and arrange for cost-effective, top-notch medical treatments abroad.

Transparency and quality are the cornerstones of our mission. We understand that choosing the right healthcare provider is a critical decision, and we've integrated the three factors that matter most to physicians when selecting or referring patients to healthcare providers: patient satisfaction, experience match, and the quality of the hospitals where our physicians practice.

Our aspiration is to lead the way in making information about physicians and hospitals more accessible and transparent. We want to empower you with the confidence to make informed healthcare choices that best suit your needs.

MedTourEasy's overarching mission is simple yet profound: to ensure that everyone, regardless of their location, time constraints, or budget, has access to high-quality healthcare. We connect patients with internationally-accredited clinics and hospitals, creating a bridge to a world of healthcare excellence.

Choose MedTourEasy for a healthier tomorrow, today. Your well-being knows no boundaries, and neither do we.

1.2 About the Project

The project, titled "Blood Donation Prognostication," addresses a critical and recurrent challenge faced by blood collection managers - the ability to accurately predict blood supply needs. Blood transfusion is a life-saving medical procedure that plays a pivotal role in healthcare, ranging from addressing injuries and surgeries to treating various illnesses and blood-related disorders. Ensuring a consistent and ample supply of blood when required is of paramount importance for healthcare professionals. The demand for blood is dynamic and fluctuates throughout the year, with notable decreases during busy holiday seasons. Accurately forecasting future blood supply levels empowers timely action, ultimately saving more lives.

Project Overview:

The project centers on the analysis of a comprehensive blood transfusion dataset and involves the following key steps:

- Data Acquisition:
  Loading the blood transfusion dataset.
  Initial inspection and exploration of the dataset to gain insights.
- Feature Selection:
  Identifying the relevant features and target columns that will be instrumental in predicting blood donation patterns.
- Data Partitioning:
  Dividing the dataset into training and testing subsets, ensuring that the model is trained on a representative portion of the data.
- Model Selection:
  Leveraging advanced automated tools such as TPOT to assist in the selection of the most appropriate predictive model.
- Model Development:
  Building the chosen predictive model that will help forecast blood donation events.
- Model Training:
  Training the predictive model using the training dataset, fine-tuning its parameters to optimize performance.
- Model Evaluation:

Rigorous assessment of the model's predictive accuracy and reliability, using appropriate evaluation metrics.

Conclusion:

In conclusion, the project, "Blood Donation Prognostication," addresses a pressing healthcare challenge by harnessing the power of data and advanced predictive modeling techniques. By accurately forecasting blood supply needs, this project contributes to the timely allocation of resources, thereby saving lives. It underscores the importance of data-driven decision-making in healthcare and stands as a beacon of hope for more efficient and effective blood supply management.

2. Methodology

2.1 Flow of the Project

The flow of the project can be seen below:

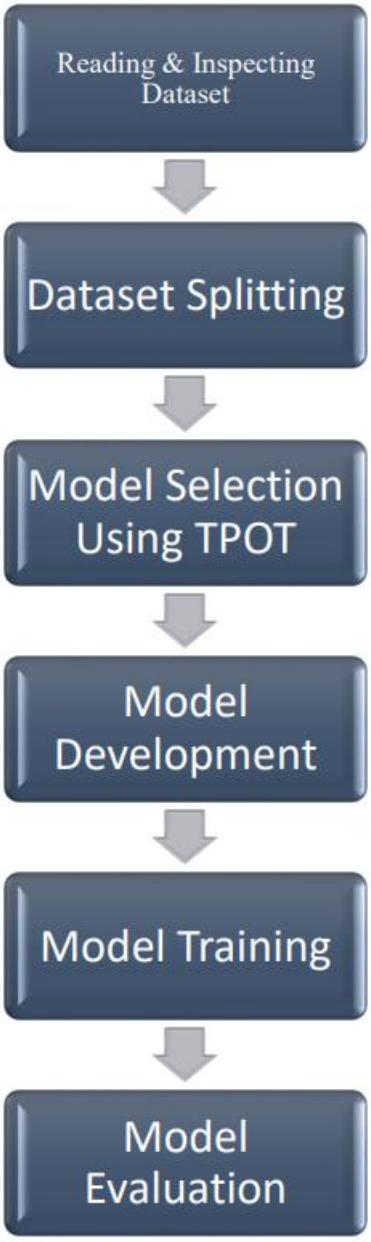

Figure 1: Flow of the Project

2.2 Language and Platform used:

Language: ***Python***

Python is a versatile and widely-used high-level programming language known for its simplicity and readability. Created by Guido van Rossum and first released in 1991, Python has gained immense popularity across various

domains, from web development and data analysis to scientific research and artificial intelligence.

Key characteristics and aspects of Python include:

Readability: Python's syntax is designed to be clear and easy to read, emphasizing code readability. This makes it an excellent choice for both beginners and experienced programmers.

Figure 2: Guido van Rossum

Versatility: Python is a general-purpose programming language, meaning it can be used for a wide range of applications, including web development, data analysis, scientific computing, automation, and more.

Large Standard Library: Python comes with a comprehensive standard library that provides modules and functions for many common tasks, reducing the need to write code from scratch and speeding up development.

Cross-Platform Compatibility: Python is available for various operating systems (Windows, macOS, Linux), making it highly portable. Code written in Python can typically run on different platforms without modification.

Interpreted Language: Python is an interpreted language, which means that you can run code directly without the need for compilation. This makes development and debugging faster.

Community Support: Python has a vibrant and active community of developers and enthusiasts who contribute to its growth. This community support is valuable for learning and problem-solving.

Open Source: Python is open-source, which means it is freely available, and anyone can contribute to its development. This open nature has contributed to Python's rapid evolution and widespread adoption.

Diverse Ecosystem: Python boasts a rich ecosystem of libraries and frameworks. For example, Django and Flask for web development, NumPy and pandas for data manipulation, TensorFlow and PyTorch for machine learning, and many more specialized tools for various domains.

Object-Oriented: Python supports object-oriented programming, facilitating the organization and management of code through classes and objects.

Dynamic Typing: Python is dynamically typed, which means you don't need to declare variable types explicitly. This can lead to more concise code and faster development.

Highly Extensible: Python can be easily extended with modules and packages written in other languages, such as C and C++, allowing developers to integrate existing code seamlessly.

Python's versatility, readability, and extensive ecosystem have made it a preferred choice for developers across industries, contributing to its status as one of the most popular programming languages in the world. Whether you're a beginner learning to code or an experienced developer working on complex projects, Python offers a robust and enjoyable programming experience.

Machine Learning:

Machine Learning (ML) is a subfield of artificial intelligence (AI) that focuses on the development of algorithms and models that enable computer systems to learn and make predictions or decisions without being explicitly programmed. It is a transformative technology that has gained immense popularity and has applications in various domains, from healthcare and finance to self-driving cars and recommendation systems.

Key concepts and aspects of machine learning include:

Learning from Data: Machine learning systems are designed to learn from data. They analyze and process large datasets to identify patterns, relationships, and trends that would be difficult or impossible for humans to discover through manual programming.

Predictive Modeling: ML algorithms build predictive models based on historical data. These models can make predictions or classifications for new, unseen data.

Supervised Learning: In supervised learning, the algorithm is trained on a labeled dataset, meaning the input data is paired with corresponding target labels or outcomes. The algorithm learns to map inputs to outputs, making it suitable for tasks like classification and regression.

Unsupervised Learning: Unsupervised learning involves working with unlabeled data. Algorithms in this category discover patterns or structure within the data, such as clustering similar data points together or reducing the dimensionality of the data.

Deep Learning: Deep learning is a subset of machine learning that focuses on neural networks with many layers (deep neural networks). It has achieved remarkable success in tasks such as image and speech recognition and natural language processing.

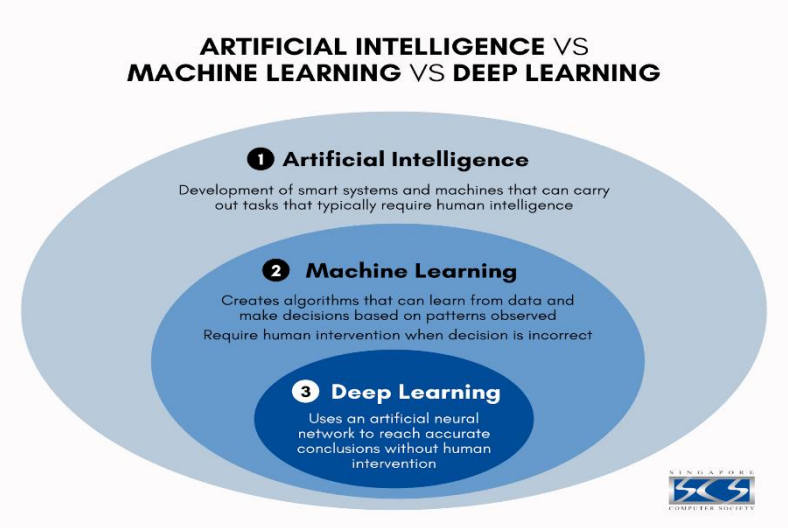

Figure 3: AI vs ML vs DL

Feature Engineering: Feature engineering involves selecting and transforming the relevant features (input variables) to improve the performance of machine learning models. It plays a crucial role in the success of many ML projects.

Model Evaluation: ML models must be rigorously evaluated using various metrics to assess their accuracy, generalization ability, and potential for overfitting or underfitting.

Applications: Machine learning has a wide range of applications, including recommendation systems (e.g., Netflix's movie recommendations), autonomous vehicles, medical diagnosis, fraud detection, natural language processing (e.g., chatbots and language translation), and more.

Continuous Learning: ML models can adapt and improve over time with new data, making them suitable for tasks that require ongoing adjustments and optimization.

Ethical Considerations: As machine learning becomes more prevalent, ethical considerations regarding bias in data, privacy, and decision-making transparency have gained significant attention.

Machine learning has revolutionized many industries by automating decision-making processes, enhancing efficiency, and enabling the development of intelligent systems. It continues to evolve rapidly, with new algorithms and techniques emerging to address increasingly complex problems. As a result, machine learning is a dynamic and exciting field with numerous opportunities for research, development, and application.

Platform: ***Jupyter Notebook***

Jupyter Notebook is a popular and powerful open-source web application used for interactive computing, data analysis, data visualization, and scientific computing. It provides an interactive environment where you can combine code execution, rich-text documentation, mathematical equations, and visualizations all in one place. Jupyter Notebook is widely employed by researchers, data scientists, educators, and developers for a variety of tasks, from prototyping code to sharing research findings and creating educational materials.

3. Implementation

3.1 Dataset Description

Blood transfusion saves lives - from replacing lost blood during major surgery or a serious injury to treating various illnesses and blood disorders. Ensuring that there's enough blood in supply whenever needed is a serious challenge for the health professionals. According to WebMD, "about 5 million Americans need a blood transfusion every year".

Our dataset is from a mobile blood donation vehicle in Taiwan. The Blood Transfusion Service Center drives to different universities and collects blood as part of a blood drive. We want to predict whether or not a donor will give blood the next time the vehicle comes to campus.
RFMTC is a variation of the RFM model. Below is a description of what each column means in our dataset:

R (Recency - months since the last donation)

F (Frequency - total number of donation)

M (Monetary - total blood donated in c.c.)

T (Time - months since the first donation)

a binary variable representing whether he/she donated blood in March 2007 (1 stands for donating blood; 0 stands for not donating blood)

3.2 Statistical Insights from the Dataset

Every dataset contains multiple hidden information which can be seen by performing statistical analysis on it. The insights found by performing statistical analysis on our dataset are the following:-

- Every column in our dataset is of numeric type.
- After specifying the features and target column, it is found that the distribution of target class in target columns are as:
    - 0 : 0.762
    - 1 : 0.238
- Due to the uneven distribution of both classes, splitting of dataset is performed using the following parameters:-
    - Test size : 0.25
    - Random State : 42

- Stratify : Target column

- The correlation matrix describing the correlation between different features through a heatmap is shown below:-

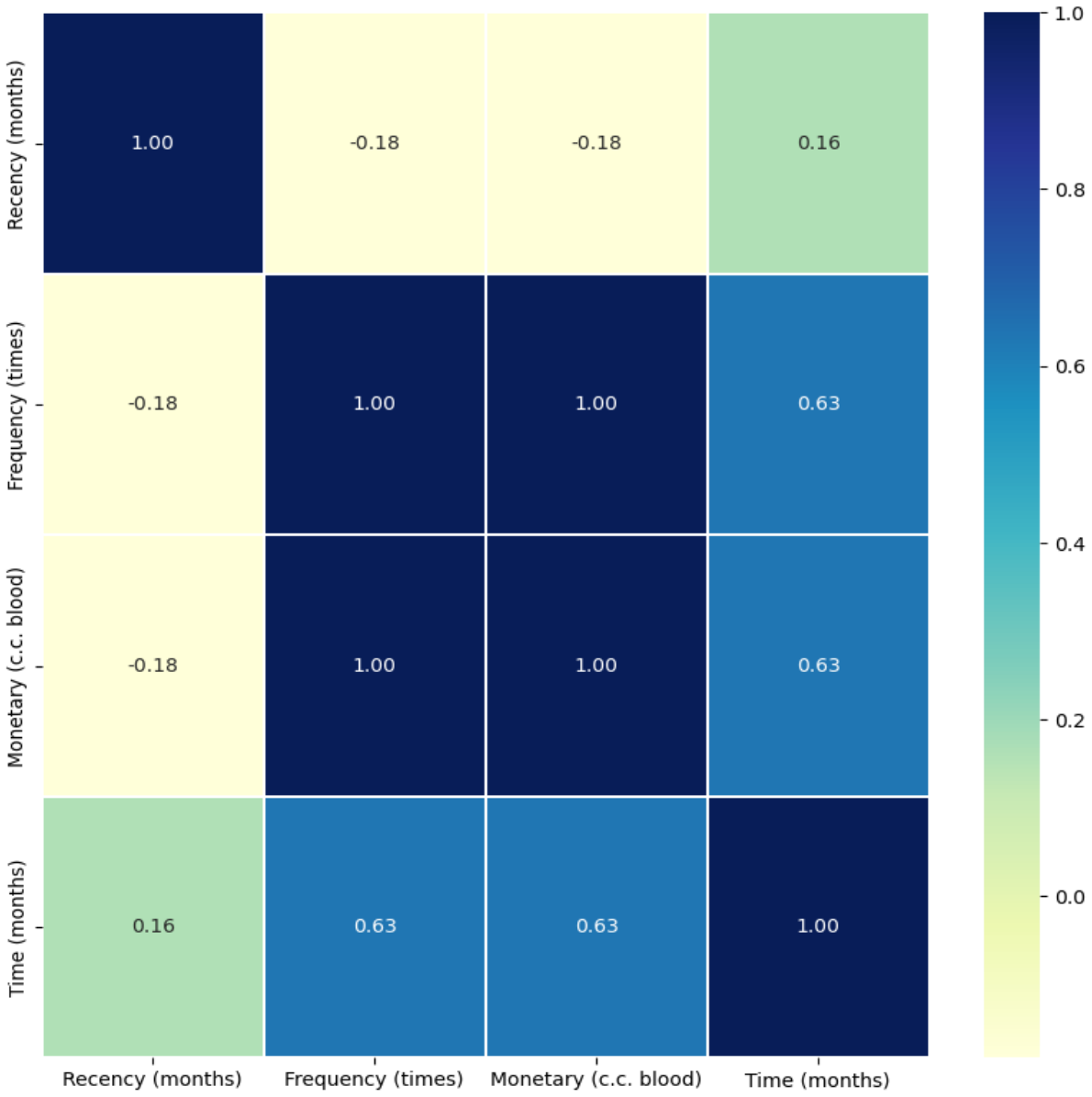

Figure 4: Dataset Correlation Graph

3.3. Model Selection and Development

To select a perfect algorithm and whole development pipeline is quite a tedious and time taking task. Therefore, we use Auto ML which defines the best pipeline and best machine learning algorithm with the best parameters.

TPOT, short for "Tree-based Pipeline Optimization Tool," is a powerful automated machine learning (AutoML) library in Python. It's designed to streamline and simplify the process of building and optimizing machine learning pipelines. TPOT is especially valuable for data scientists, machine learning practitioners, and researchers who want to automate the tedious and time-consuming aspects of model selection and hyperparameter tuning.

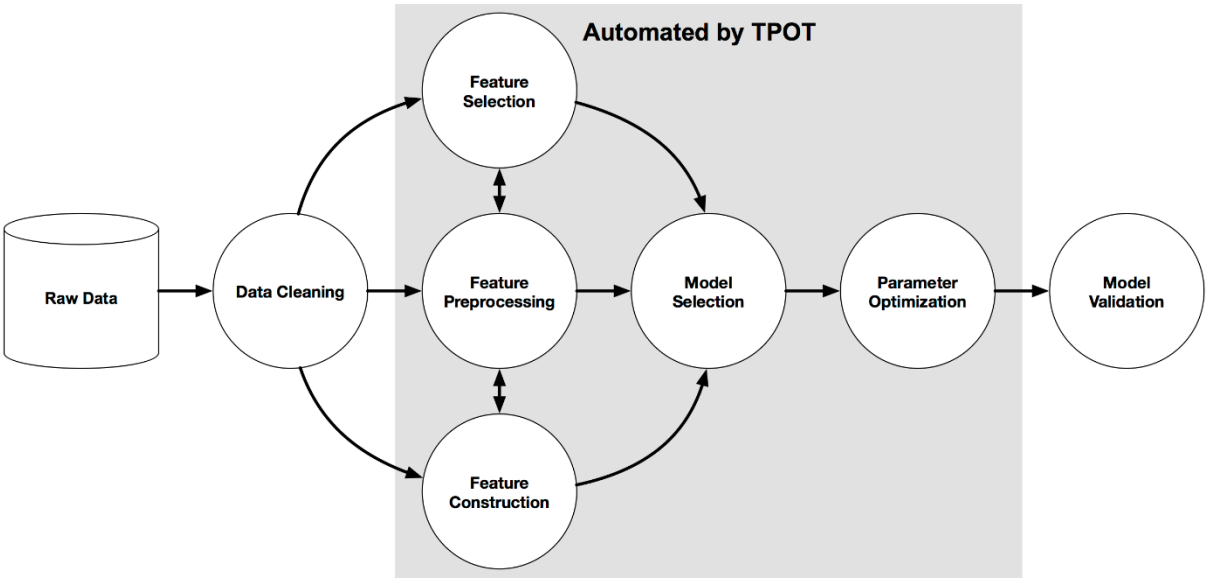

Figure 5: TPOT

TPOT will automatically explore hundreds of possible pipelines to find the best one for our dataset. Note, the outcome of this search will be a scikit-learn pipeline, meaning it will include any pre-processing steps as well as the model.

We are using TPOT to help us zero in on one model that we can then explore and optimize further.

TPOT picked “Logistic Regression” as the best model for our dataset with no pre-processing steps, giving us the AUC score of 0.7858. Therefore, I have used Logistic regression algorithm for model development.

3.4 Model Training and Evaluation

One of the assumptions for linear regression models is that the data and the features we are giving it are related in a linear fashion, or can be measured with a linear distance metric. If a feature in our dataset has a high variance that's an order of magnitude or greater than the other features, this could impact the model's ability to learn from other features in the dataset. Correcting for high variance is called normalization. It is one of the possible transformations we do before training a model. Monetary (c.c. blood)'s variance was very high in comparison to any other column in the dataset. This means that, unless accounted for, this feature might get more weight by the model (i.e., be seen as more important) than any other feature. One way to correct for high variance was to use log normalization.

Finally, the logistic regression model is trained with the following parameters:

- Solver = lbfgs
- Random State = 42

The model evaluation is performed by checking AUC Score of the model which is 0.7891.

4. Conclusion and Future Scope

The demand for blood exhibits dynamic fluctuations, notably during peak holiday seasons when blood donations tend to decrease. The ability to accurately forecast future blood supply needs is paramount as it allows for proactive measures to be taken, ultimately preserving more lives.

In the pursuit of enhancing prediction accuracy, we embarked on a journey into automated model selection using TPOT, a cutting-edge machine learning tool. The fruits of our labor were indeed promising, with an achieved AUC score of 0.7891. This performance surpasses the baseline scenario of simply choosing 0 as the prediction, a strategy that aligns with the target incidence, thereby suggesting a 78% success rate.

Not content with resting on our laurels, we delved deeper into refining our approach. By applying log normalization to our training data, we were able to elevate the AUC score further, boasting a remarkable 0.5% improvement. In the realm of machine learning, even seemingly modest enhancements in accuracy can hold profound significance, contingent upon the specific context and objectives.

Another compelling facet of our strategy is the utilization of a logistic regression model, a choice characterized by its interpretability. This affords us the invaluable opportunity to dissect and comprehend the extent to which the variance in the response variable, i.e., the target, can be elucidated by the other variables within our dataset. In essence, this lends transparency and insight into the intricate dynamics at play, providing a deeper understanding of the factors influencing our predictions.

References

The following websites have been referred for input data and statistics:-

- https://www.webmd.com/a-to-z-guides/blood-transfusion-what-toknow#1
- https://www.kjrh.com/news/local-news/red-cross-in-blood-donation-crisis
- https://www.ncbi.nlm.nih.gov/books/NBK310569/
- http://epistasislab.github.io/tpot/

The following websites have been referred for coding part:-

- https://www.python.org/
- https://github.com/perborgen/LogisticRegression
- http://epistasislab.github.io/tpot/