NANDINI GANTAYAT

Statistical & Machine Learning Individual project
Prof.Minh Phan

IESEG School of Management
MSc in Big Data Analytics for Business

Project Objective

(1) the understanding of machine learning mechanism
(2) the ability to setup a machine learning pipeline.

Bank Telemarketing Outcome Prediction
Predict if the client will subscribe a term deposit after a bank telemarketing campaign

MACHINE LEARNING MODELS

1. Logistic regression
2. Random forest
3. K-nearest neighbours(KNN)
4. Gradient Boosting(GBM)
5. Neural network

Logistic regression

Logistic regression is a statistical model that is used to predict binary outcomes or so it is well suited for. It is a type of linear model that estimates the probability of an event occurring based on one or more predictor variables. The objective of logistic regression is to find the best-fitting model that maximizes the likelihood of observing the data given the model parameters.

The logistic regression model is characterized by its sigmoid-shaped curve, which represents the probability of the binary outcome as a function of the predictor variables. The output of the model is a probability value between 0 and 1, which is then converted into a binary prediction using a threshold value.

There are two main types of logistic regression: binary logistic regression and multinomial logistic regression. Binary logistic regression is used when the outcome variable has two categories, while multinomial logistic regression is used when the outcome variable has more than two categories.

The advantages of logistic regression include its simplicity, interpretability, and ability to handle both categorical and continuous predictor variables. It is also a robust model that can handle outliers and missing data. Additionally, logistic regression can be used for feature selection and variable reduction. However, there are also some disadvantages to using logistic regression. For example, it assumes that the relationship between the predictor variables and the outcome variable is linear, and it may not perform well when the data is not linearly separable. Additionally, logistic regression may not work well for complex datasets with many predictor variables.

I believe due to its simplicity a lot of crowd neglects it and opts for a complex model. But the base needs to be always robust. Sometimes a simple model solves problems that a complex would be unable to.

Random forest

Random forest is a supervised machine learning algorithm that is used for classification and regression tasks. It is an ensemble learning method that combines multiple decision trees to create a more accurate and robust model. The objective of random forest is to create a forest of decision trees that collectively make predictions based on a set of input features. It ensemble technique helps it to negate out model weaknesses

It is also resistant to overfitting, which occurs when a model is too complex and captures noise in the training data instead of the underlying patterns. With all the projects I worked with tree based models never seem to disappoint me.

The random forest algorithm works by randomly selecting subsets of the input features and a subset of the training data to build each decision tree. This process is repeated multiple times to create an ensemble of decision trees, where each tree is trained on a different subset of features and data. The final prediction is made by aggregating the predictions of all the trees in the forest.

There are two main types of random forest: classification random forest and regression random forest. Classification random forest is used for predicting categorical outcomes, while regression random forest is used for predicting continuous outcomes.

The advantages of random forest include its ability to handle high-dimensional data, non-linear relationships, and interactions between variables. It is also robust to noise and can handle missing data. Random forest can be used for feature selection and variable importance analysis.

However, there are also some disadvantages to using random forest. For example, it may not perform well when the data has many irrelevant features, as it can become computationally expensive and may overfit.

K-nearest neighbours

KNN is a supervised machine learning algorithm used for classification and regression tasks. It is a non-parametric method that does not make any assumptions about the distribution of the data. Instead, it works by finding the k nearest neighbors to a given data point and using their labels or values to make a prediction.

The KNN algorithm is characterized by its simplicity and flexibility. It can work well with both linear and non-linear relationships in the data and can handle multi-class classification problems.

The KNN algorithm works by calculating the distance between the test data point and all the training data points. It then selects the k nearest neighbors and uses their labels or values to make a prediction. The value of k is a hyperparameter that needs to be tuned based on the dataset.

There are two main types of KNN: classification KNN and regression KNN. Classification KNN is used when the target variable is categorical, while regression KNN is used when the target variable is continuous. If I have a non linear data I would opt KNN as my first preference

The advantages of KNN include its simplicity, flexibility, and ability to work well with both linear and non-linear relationships in the data. However, there are also some disadvantages to using KNN. For example, it can be sensitive to the choice of k, and the distance metric used to calculate the distance between data points can also have a significant impact on the results.

Gradient boosting

GBT is a supervised machine learning algorithm that is used for classification and regression tasks. It is an ensemble learning method that combines multiple decision trees to create a more accurate and robust model. The objective of GBT is to iteratively add decision trees to the model, with each tree attempting to correct the errors made by the previous trees.

The GBT algorithm works by first creating a decision tree with a single node. It then calculates the residuals of the predicted values and fits a new decision tree to the residuals. This process is repeated multiple times, with each subsequent tree attempting to correct the errors made by the previous trees. The final prediction is made by aggregating the predictions of all the trees in the ensemble.

There are two main types of GBT: gradient boosting for classification (GBM) and gradient boosting for regression (GBR). GBM is used when the target variable is categorical, while GBR is used when the target variable is continuous.

The advantages of GBT include its ability to handle high-dimensional and non-linear data, its resistance to overfitting, and its ability to capture complex relationships between variables. GBT can be used for feature selection and variable importance analysis. However, there are also some disadvantages to using GBT. For example, it can be computationally expensive and slow to train, especially for large datasets.

Neural networks

A neural network is a type of machine learning algorithm that is modeled after the structure and function of the human brain. It consists of layers of interconnected nodes, or neurons, that process information and make predictions. Neural networks can be used for a variety of tasks, including image and speech recognition, natural language processing, and predictive modeling.

The basic unit of a neural network is a neuron, which receives input from other neurons and produces an output based on its activation function. Neurons are organized into layers, with each layer performing a different type of computation. The first layer is typically the input layer, which receives the raw data. The last layer is the output layer, which produces the final predictions. The most common type of neural network is the feedforward neural network, which consists of input, hidden, and output layers. The input layer receives the raw data, and the hidden layers perform computations on that data before passing it to the output layer. There are also other types of neural networks, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), which are designed for specific tasks.

Neural networks are trained using a process called backpropagation. During training, the network is presented with a set of labeled examples, and the weights between the neurons are adjusted to minimize the error between the predicted and actual outputs. This process is repeated until the network produces accurate predictions on new, unseen data.

The advantages of neural networks include their ability to learn complex, non-linear relationships in the data, their flexibility in handling different types of data, and their ability to generalize to new, unseen data. Neural networks can also handle missing data and noisy data. However, there are also some disadvantages to using neural networks. They can be computationally expensive and require a large amount of data to train. Neural networks can also be difficult to interpret, and it can be challenging to diagnose and fix problems with the model. No matter their disadvantages for text and speech classification they are always superior

Data Overview

Bank Marketing Data Set: The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns are based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

The data has 21 columns and 20,000 rows in total. This includes 20 independent variables and 1 dependent variable ‘subscribed’. The target value distribution is skewed

Data Pre-Processing

1) The data has no NA’s.

2)The columns Age, campaign, pdays and previous had outliers.

3) The categorical variables were encoded using Ordinal Encoder.

4) Data was split in train and test before any feature engineering with the test size of 20% using stratified split.

Feature Engineering

1) Polynomial terms of degree 2 and 3 were added to numerical variables.

2) Categorical variables were mapped using decision tree based re-mapping and were grouped based on the last node as the category.

3) Numerical variables were categorized and grouped in buckets and then binned in categories.

4) Weight of Evidence was applied to the categorical variables and were dummy encoded based on the incident rates per category.

5) mutual information was used to gain information about variables

6) Total number of features engendered using the above techniques was 788.

7) All new featured with low or no variance and duplicate values were dropped leading to 411 variables.

Feature selection

Correlation: Correlation is a statistical measure that describes the strength and direction of the relationship between two variables. It can range from -1 to 1, with -1 indicating a perfect negative correlation, 0 indicating no correlation, and 1 indicating a perfect positive correlation.

Fisher's score: Fisher's score is a statistical measure that is used to evaluate the importance of features in a classification problem. It is based on the difference between the mean values of the feature for each class and the variance of the feature.

RFE - Recursive Feature Selection: It works by recursively removing features and building a model on the remaining features until the desired number of features is reached.

Boruta: Boruta is a feature selection technique that is designed to handle high-dimensional data with correlated features. It works by comparing the importance of each feature to a set of shadow features, which are created by permuting the values of the original features. Features that are more important than the shadow features are considered significant, while features that are less important are considered unimportant.

t-SNE: t-SNE (t-distributed stochastic neighbor embedding) is a dimensionality reduction technique that is used to visualize high-dimensional data in two or three dimensions. It works by converting the high-dimensional data into a lower-dimensional space while preserving the relationships between the data points.

Model evaluation and performance

GBM model with fisher score and Boruta feature selection performs the best whereas all the models perform the worst with correlation features selection

	Model_w_fisher_feat	Train Accuracy	Test Accuracy	Train AUC	Test AUC
0	Logistic Regression	0.900750	0.89625	0.788169	0.786198
1	Random Forest	0.926937	0.89250	0.858730	0.762459
2	KNN	0.908188	0.89400	0.787803	0.732836
3	GBM	0.906250	0.89725	0.806929	0.794824
4	NN	0.903687	0.89625	0.797966	0.785475

	Model_w_BORUTA_feat	Train Accuracy	Test Accuracy	Train AUC	Test AUC
0	Logistic Regression	0.901813	0.89550	0.789940	0.794350
1	Random Forest	0.951187	0.88275	0.953717	0.728244
2	KNN	0.912375	0.88475	0.873106	0.702208
3	GBM	0.907500	0.89450	0.805487	0.795339
4	NN	0.905500	0.89675	0.796973	0.787564

References

https://www.linkedin.com/pulse/correlation-vs-causation-muhammad-umair-ali

https://learnopencv.com/t-sne-t-distributed-stochastic-neighbor-embedding-explained/

ISLRV BOOK