**Comparing Classifiers**

In this practical application, our goal is to compare the performance of the classifiers we encountered in this section, namely K Nearest Neighbor, Logistic Regression, Decision Trees, and Support Vector Machines. We will utilize a dataset related to marketing bank products over the telephone.

We used industry standard process, named CRISP-DM framework, to work through a data problem.

1. **Business Understanding**

The objective of this project is to compare different classifiers (k-nearest neighbors, logistic regression, decision trees, and support vector machines) using the bank marketing dataset from the UCI Machine Learning Repository. The goal is to determine the most accurate and reliable model for predicting the success of a marketing campaign for a bank.

The dataset contains information about client attributes, campaign details, and outcome variables related to bank marketing campaigns. By evaluating the performance of different classifiers on this dataset, the objective is to assess their accuracy, precision, recall, and F1-score in predicting whether a client will subscribe to a term deposit or not.

The aim is to identify the classifier that performs best in terms of these metrics, which will provide valuable insights for the bank's marketing campaign.

Thank you, UCI Machine Learning repository, for allowing us to use your dataset <https://archive.ics.uci.edu/dataset/222/bank+marketing> .

1. **Data Understanding: Exploratory Data Analysis (EDA)**

* Load and Read Data Set
* Understand Features and Datatype
* Analyse Data and Statistics

About Dataset

The dataset contains total 21 features/variables including target variable named y. Here's brief about each feature:

Input variables:  
# bank client data:  
1 - age (numeric)  
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')  
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)  
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')  
5 - default: has credit in default? (categorical: 'no','yes','unknown')  
6 - housing: has housing loan? (categorical: 'no','yes','unknown')  
7 - loan: has personal loan? (categorical: 'no','yes','unknown')  
# related with the last contact of the current campaign:  
8 - contact: contact communication type (categorical: 'cellular','telephone')  
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')  
10 - day\_of\_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')  
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.  
# other attributes:  
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)  
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)  
14 - previous: number of contacts performed before this campaign and for this client (numeric)  
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')  
# social and economic context attributes  
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)  
17 - cons.price.idx: consumer price index - monthly indicator (numeric)  
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)  
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)  
20 - nr.employed: number of employees - quarterly indicator (numeric)  
  
Output variable (desired target):  
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

1. **Data Preparation**

* Drop features & duplicates
* Check and Handle Missing Values
* Feature Encoding
* Visualize categorical, target, feature correlation

1. **Modeling**

We have analysed the data against various classification models using default hyper parameters and optimum hyperparameters through GridSearchCV technique. The classifier models used in current applications are:

* Logistic Regression Model
* KNN Model
* Decision Tree Model
* SVM Model

1. **Evaluation**

In order to determine the most suitable model, we have performed the following steps:

* Model Comparison i.e. Accuracy (Test/Train), Fit Time etc.
* Visualize Accuracy vs Time
* Best Params
* Feature Importance / Coefficient
* ROC Curve
* Visualize Classification Report i.e. Accuracy, Precision, Recall etc.
* Visualize Confusion Matrix

1. **Deployment**

* Save the trained model to a file for future use.
* Load the saved model and use it to make predictions on new data.

**Key Highlights**

* Decision Tree model performs the best in terms of accuracy and training time (8 seconds).
* Decision Tree model has the best accuracy (0.92), recall, and F1-Score metrics among all four models.
* Logistic Regression model has the best precision and ROC-AUC Curve metrics among all four models.
* SVM model (SVC classifier) has the highest training time (1000+ seconds) and is computationally expensive.
* Decision Tree: Optimized parameter model has the 0.02 more accuracy compare to default parameter model.
* Bank Marketing dataset is impbalanced because distribution is not similar within dataset.
* The highest positive correlation between numerical independent variables is between "emp.var.rate" and "euribor3m" with a value of 0.97.
* The highest negative correlation between numerical independent variables is between "pdays" and "previous" with a value of -0.59.
* Decision Tree model predicted the highest True Positives, while SVM (SVC Classifier) predicted the highest True Negatives, followed by the Logistic Regression model on the Test Data.
* The "Day\_of\_Week" feature has a similar success rate among all categorical values, so it is not an important feature for the given model dataset.
* The "nr.employed" & "euribor3m" are higher importance features since they have higher influence towards model predictions.
* The "month", "cons.price.idx" and "duration" features have higher coefficient and they contribute higher towards predicting the target variable.
* Overall, the Decision Tree model is the best performing model among all four models followed by Logistic Regression model.

**Next steps and Recommendations**

* Evaluate and identify non-important categorical features in a group (with other feature variable) rather than single to determine whether there exists combination of categorical features which can be discarded.
* Continue to add samples from new marketing campaign and evaluate the different classifier models.
* Imbalanced dataset: Since the bank marketing dataset is imbalanced, consider using techniques such as resampling, SMOTE, or cost-sensitive learning to address the class imbalance and improve model performance.
* Feature importance: Focus on the "nr.employed", "euribor3m", "month", "cons.price.idx", and "duration" features, as they have higher coefficients or importance towards predicting the target variable. Consider further analyzing these features and their relationships with the target variable to gain deeper insights.
* Training time and computational expense: Given that the SVM model (SVC classifier) has the highest training time and is computationally expensive, consider optimizing the model or exploring alternative models that can provide similar performance with lower training time.
* Performance metrics: While the Decision Tree model performs the best in terms of accuracy, recall, and F1-Score, and the Logistic Regression model has the best precision and ROC-AUC Curve metrics, it is important to consider the specific requirements of the problem at hand. Evaluate the trade-offs between different metrics and select the model that aligns best with the desired outcome.
* Correlation analysis: Take note of the highest positive correlation between "emp.var.rate" and "euribor3m" and the highest negative correlation between "pdays" and "previous". These insights can help in understanding the relationships between variables and potentially identifying multicollinearity issues.
* Overall, continue to refine and optimize the Decision Tree model and consider the recommendations mentioned above to further enhance model performance and interpretability.