Task 3: Data Analysis Lessons Learned

Its important to understand the real business problem which we are trying to solve. Hence, carefully review and understand the requirements.

1. **Obtaining, cleaning and Importing Data**:
   * Data preparation - – finding out about the data available to understand – no. of attributes, type of attributes, missing values and anything unusual which needs to addressed prior to the detail analysis e.g. special characters.
   * Data Import – Use the appropriate operator to import data in RapidMiner depending on the format of data available.
2. **Initial Exploration of Data:** Understand your data to help determine the type of algorithms you choose for modeling
   * Computing summary statistics and understanding the structure of the data serve to inform potential changes needed in preprocessing
   * **Data Visualization – Use of** a histogram visualize the various distributions of each attribute in the data and use scatter plots to compare the relationships between any two features.
3. **Preprocessing:** 
   * Decide on handling of missing values – remove, for numerical data – use mean, mode or median values
   * Assign the right data types – identify if there are need to correct or change the data attribute types – Binomial/polynomial
4. **Feature Selection and Feature Engineering:**
   * Deep investigation of the relationships between all the independent variables and the dependent variable.
     + Importance of identification of dependent variable based on business problem you are trying to solve.
5. **Modeling and Optimization:**
   * Consider the causal relationship between the dependent variables and independent variables as proposed by the user, and endeavor to capture the degree and nature of dependence between the variables
   * Decision Tree – is versatile tool which can be used to solve both classification and regression problems. Decision trees can be used in both data mining and machine learning which involves training a model and testing it
   * **Model selection** – the application of one or more algorithms to understand patterns within data sets and to make predictions about future outcomes.
   * Cross Validation –
   * Used to estimate performance of the algorithm and help choose between models.
   * The algorithm is built using training data and the validated on the testing data
   * The training process is generally iterative and that part of this process involves revisiting things like data pre-processing and feature engineering to get the best results out of your trained models.
     + Models - Select model based on problem to addressed - classification vs. Regression.
     + Accessing Model Performance: assess the performance of the model and the accuracy of its predictions.
   * Choose the model that produces the best results
     + Highest accuracy.
     + Least level of comparative error and the highest R-Squared
6. **Making Predictions:** Applying an optimized model to a test set or unseen data to provide predictions that can help solve business or engineering issues.
7. **Reporting Your Findings:** The best analysis is only as good as how well it is reported. Always consider your audience’s level of technical background when developing reports and focus on answering the key stakeholder questions you set out to investigate.