# Decision Tree and Random Forest

**Instructions:**

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

**Name:RAKESH SETHU NP Batch ID: 11052022\_7.30PM**

**Topic: Decision Tree and Random Forest**

**Hints:**

1. **Business Problem**
   1. **What is the business objective?**
   2. **Are there any constraints?**
2. **Work on each feature of the dataset to create a data dictionary as displayed in the below image:**



**2.1 Make a table as shown above and provide information about the features such as its data type and its relevance to the model building. And if not relevant, provide reasons and a description of the feature.**

1. **Data Pre-processing**

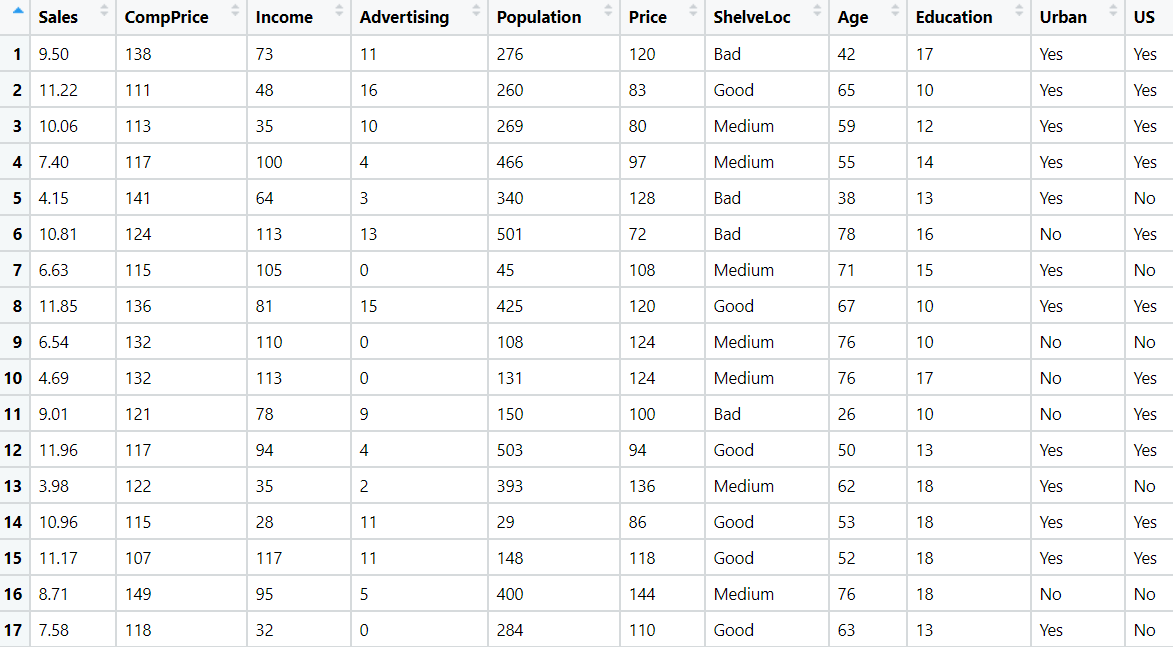
**3.1 Data Cleaning, Feature Engineering, etc.**

1. **Exploratory Data Analysis (EDA):**
   1. **Summary.**
   2. **Univariate analysis.**
   3. **Bivariate analysis.**
2. **Model Building**
   1. **Build the model on the scaled data (try multiple options).**
   2. **Perform Decision Tree and Random Forest on the given datasets.**
   3. **Train and Test the data and perform cross validation techniques, compare accuracies, precision and recall and explain about them.**
   4. **Briefly explain the model output in the documentation.**

1. **Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?**

**Problem Statements:**

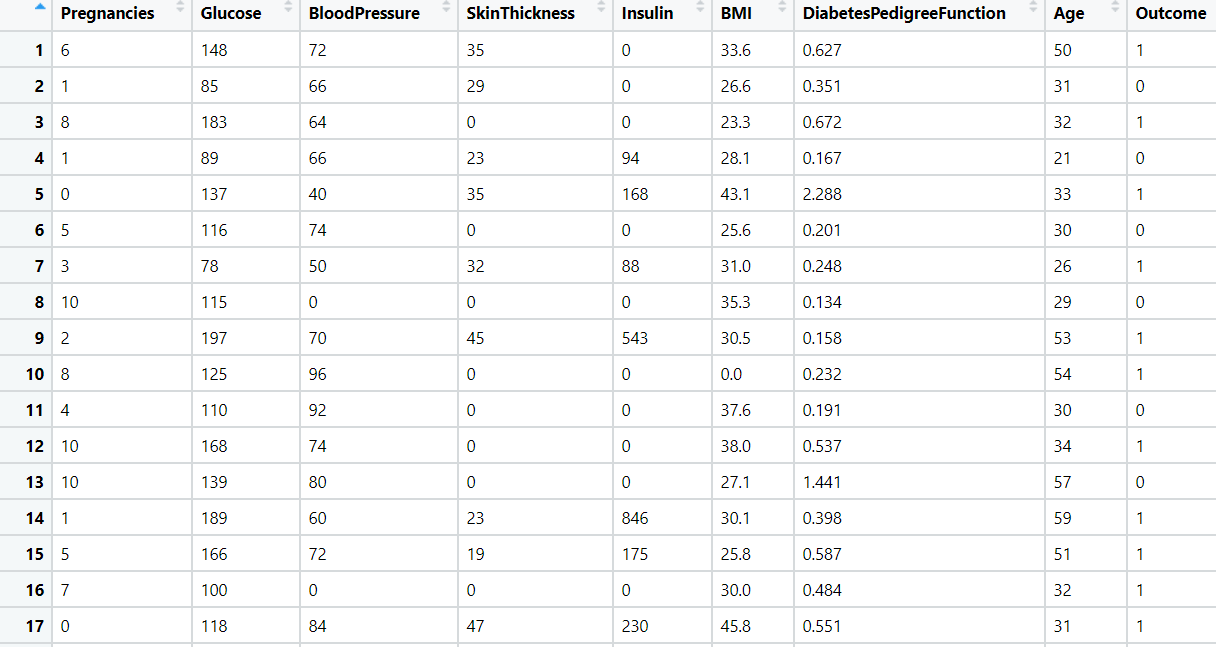
1. A cloth manufacturing company is interested to know about the different attributes contributing to high sales. Build a decision tree & random forest model with Sales as target variable (first convert it into categorical variable).



1. **Business Problem**
   1. **Objective: -** Maximize making profit
   2. **Constraints (if any): -** restriction or limitations on decision variable

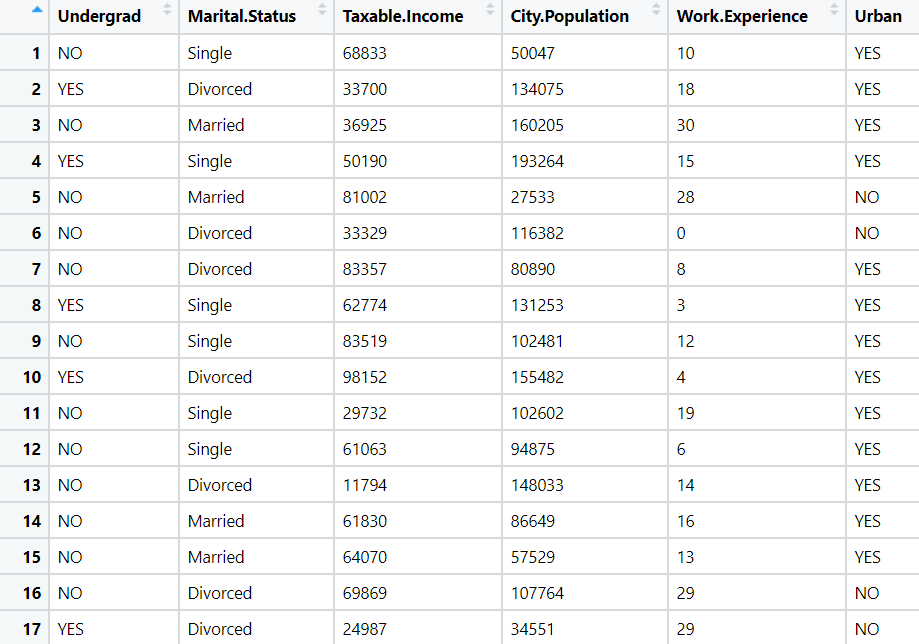
|  |  |  |  |
| --- | --- | --- | --- |
| **Name of feature** | **description** | **type** | **Relevance** |
| Sales | no of sales in company | qualitative, continues | relevant |
| Comp Price | company price for product | qualitative, discrete | relevant |
| Income | income of company | qualitative, continues | relevant |
| Advertising | Advertising given by company | qualitative, discrete | irrelevant |
| Population | no of population | qualitative, discrete | irrelevant |
| Price | price of product | quantitative discrete | relevant |
| Shelve Loc | quality of shelving location | qualitative, discrete | irrelevant |
| Age | no of age | Quantitative, discrete | irrelevant |
| Education | how many are educated | Quantitative, discrete | irrelevant |
| Urban | located in urban or not | Quantitative, discrete | irrelevant |
| US | located in us or not | Quantitative, discrete | irrelevant |

1. Divide the diabetes data into train and test datasets and build a Random Forest and Decision Tree model with Outcome as the output variable.



|  |  |  |  |
| --- | --- | --- | --- |
| **Name of feature** | **description** | **type** | **Relevance** |
| Number of times pregnant | No of times Pregnancies | Quantitative, discrete | relevant, it provides useful information |
| Plasma glucose concentration | Blood glucose levels in your body, | Quantitative, continues | relevant |
| Diastolic blood pressure | Blood Pressure in your body | Quantitative, discrete | relevant |
| Triceps skin fold thickness | Thickness of skin | qualitative ,discrete | irrelevant |
| 2-Hour serum insulin | insulin measurement | qualitative, continues | relevant |
| Body mass index | BMI rate | quantitative, continues | relevant |
| Diabetes pedigree function | Diabetes Pedigree Function | qualitative, continues | irrelevant |
| Age (years) | age of individual person | Quantitative, discrete | irrelevant |
| Class variable | over all outcome | Quantitative, discrete | relevant |





1. Build a Decision Tree & Random Forest model on the fraud data. Treat those who have taxable\_income <= 30000 as Risky and others as Good (discretize the taxable income column).



**Problem3:**

**Business Objective:** Minimize fraud.

**Business Constraint**: Maximize Convenience to customers.

|  |  |  |  |
| --- | --- | --- | --- |
| Name of Feature | Description | Type | Relevance |
| Undergrade | Inferior to standard grade | Categorical | Irrelevant, it does not provide useful information |
| Marital Status | Status of an individual person | Nominal | It is not useful |
| Taxable Income | Tax for individual person | Continuous | It provides useful information for calculations |
| City Population | Population of city | Discrete | It provides useful information |
| Work Experience | Years of experience | Continuous | It’s a useful information |
| Urban | Characteristics of city | Category | Does not provide useful information |

1. In the recruitment domain, HR faces the challenge of predicting if the candidate is faking their salary or not. For example, a candidate claims to have 5 years of experience and earns 70,000 per month working as a regional manager. The candidate expects more money than his previous CTC. We need a way to verify their claims (is 70,000 a month working as a regional manager with an experience of 5 years a genuine claim or does he/she make less than that?) Build a Decision Tree and Random Forest model with monthly income as the target variable.

A screenshot of a cell phone

Description automatically generated

**Problem 4:**

**Business Objective:** Minimize Fraud Candidate.

**Business Constraint**: Maximize the Profit.

|  |  |  |  |
| --- | --- | --- | --- |
| Name of Feature | Description | Type | Relevance |
| Position of employee | Position of person in an organization | Categorical | It is not providing useful information |
| No of years of experience | Experience year of a person | Continuous | It provides useful information |
| Monthly income of employee | Income of employee | Continuous | It provides useful information |