Lecture Notes

Lecture Notes - Model Selection Practical Considerations

With so many algorithms to choose from, which one should you really opt for? You could apply all the models and compare their results; but is it always feasible to apply all the models? You won't always have enough time to try all the available options. More importantly, it is much more helpful to be able to identify some guiding principles behind the choice of models than to use the hit-and-trial approach.

Understanding the Business Problem

An ecommerce company has decided to learn more about its online consumers. More specifically, it wants to determine the gender of each consumer and the age group to which they belong. In terms of gender, a consumer can either be male or female. The age groups they fall into can be as follows:

1. Young adults (21-26 years old)
2. Family person (26-35 years old)
3. Middle aged (25-45 years old)
4. Senior people (45+ years aged)

In total, there will be eight categories as shown below:

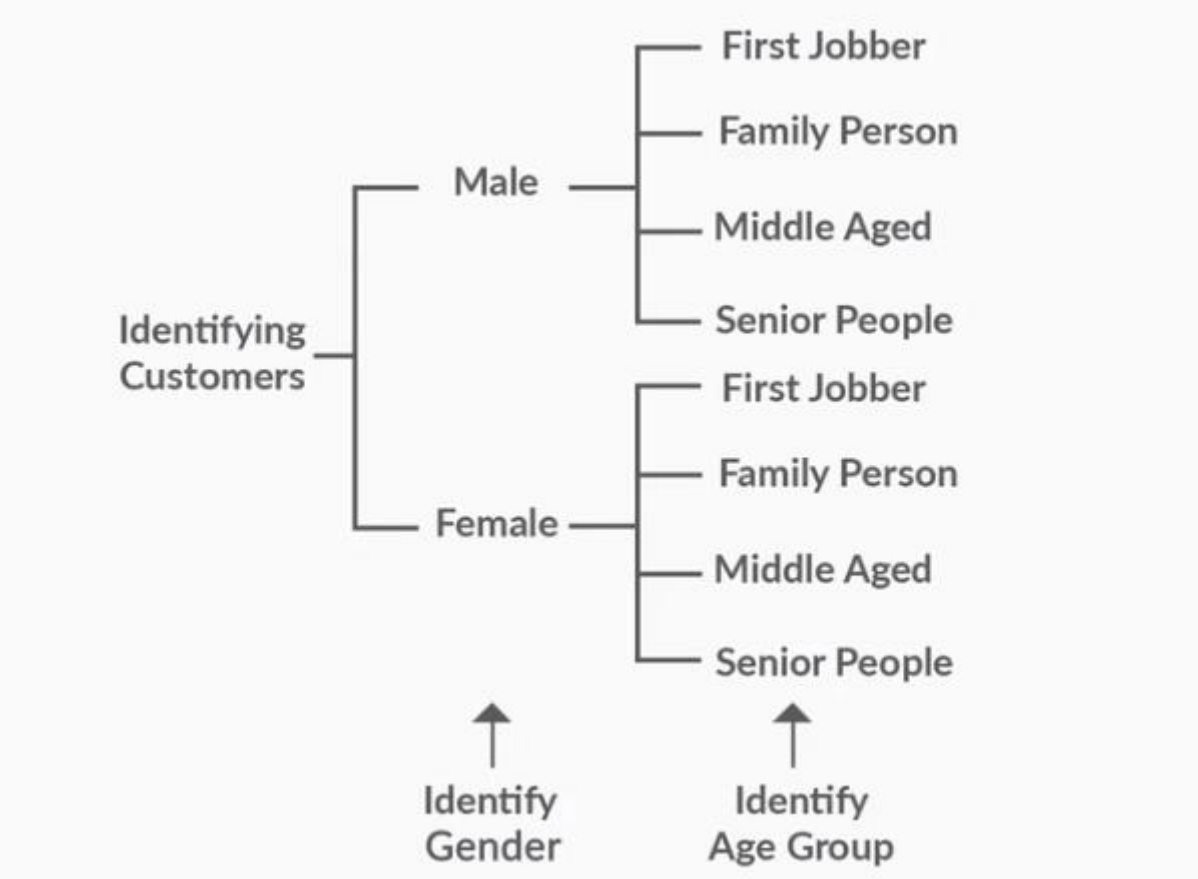

Figure 1- Identifying consumers

As shown in the image, this is a two-step process that involves **identifying the gender** and **identifying the age group** within each gender group.

The process of **identifying** and learning about the consumers will **benefit** the ecommerce company in the following ways:

1. It will help the company provide **better recommendations** to the customers, based on their gender and age group.
2. It will help it have a **higher conversion rate of products** that go from the cart to the final billing section.
3. It will also help the company **provide a better consumer experience**.

Comprehension - Logistic Regression

The **equation** for a **logistic regression** model is

$$P = \frac{1}{1+e^{-(\beta_0+\beta_1 x1+\beta_2 x2)}}$$

In the context of the business problem that you are going to solve, P denotes the probability of a consumer being male (y = 1),
x_1 is the attribute: time of the day,
x_2 is the attribute: ratio of items bought/items added to the cart, and
β_0 is the intercept term, while β_1 and β_2 are the coefficients of the attributes.

Remember that the equation given above can be rewritten as the **log odds** equation

$$ln(\frac{P}{1-P}) = \beta_0 + \beta_1 x1 + \beta_2 x2$$

In this equation, the term $\frac{P}{1-P}$ is known as the **odds**. Here, the odds indicate the chances of a consumer being male (P) as a proportion of the chances of the consumer being female (1-P).

The right-hand side of the log odds equation is used to interpret the **decision boundary** of a logistic regression model. The gender of the person can be determined using a **threshold value**, *t*. Remember that while modelling a logistic regression model, you chose a cutoff value, *c*, say, 0.5. If P > c, then the predicted output is 1; otherwise, it is 0. You can calculate *t* using *c* by substituting the value of *c* in the following equation:

$$t = ln(\frac{c}{1-c})$$

If $\beta_0 + \beta_1 x_1 + \beta_2 x_2 > t$, then the consumer is male (y =1); otherwise, the consumer is female (y =0).

Comparing Different Machine Learning Models

You used logistic regression, decision trees, and support vector machines to identify the gender of the consumers. The distribution of the consumers is shown below:

Figure 2 - Distribution of ecommerce consumers

The logistic regression boundary is shown below.

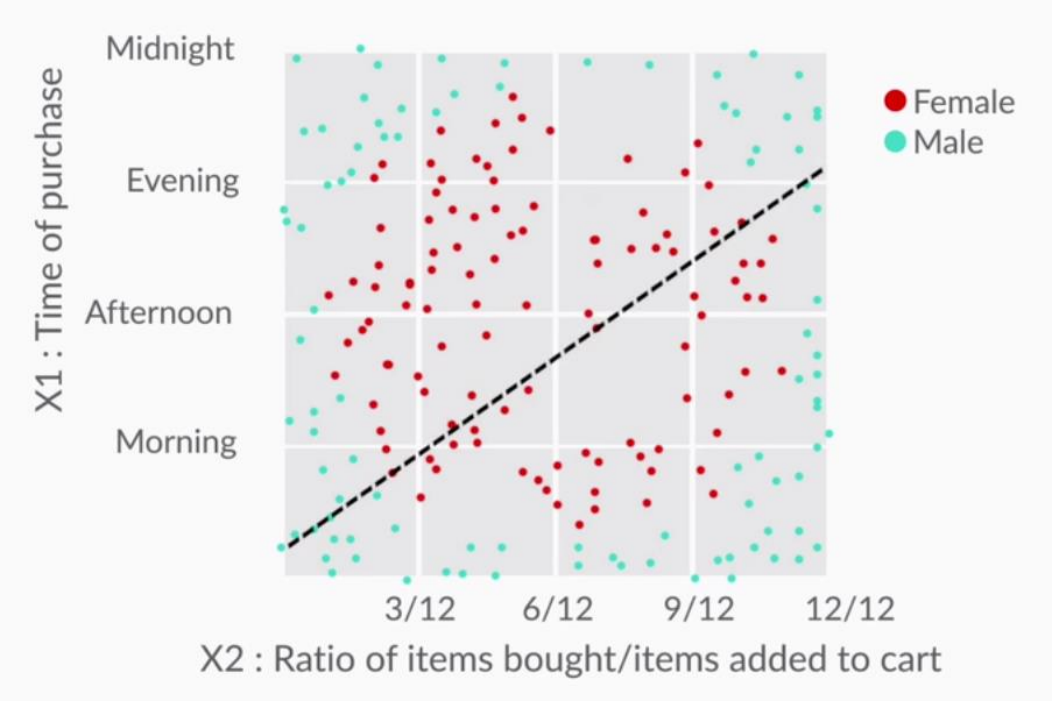

Figure 3 - Logistic regression decision boundary

The linear decision boundary of logistic regression is clearly not able to separate the two classes in this case.
Next, you saw the decision boundary of two decision trees: a simple decision tree and a complex decision tree. The decision boundary of the simple decision tree — the one with three linear boundaries — is shown below:

Figure 4 - Decision boundary of a simple decision tree

Now let's look at the decision boundary of the complex decision tree: the one with 12 decision boundaries.

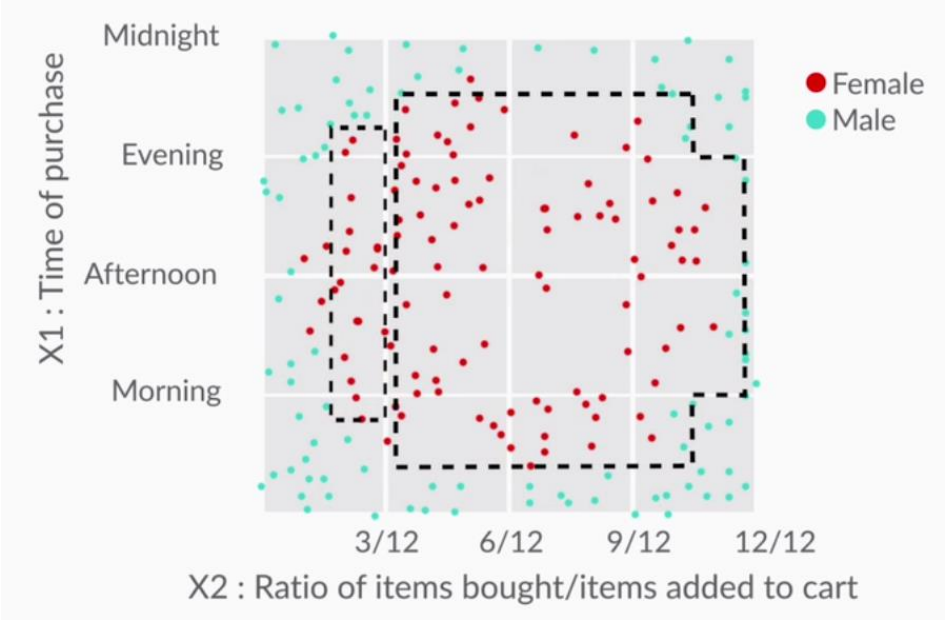

Figure 5 - Decision boundary of a complex decision tree

The decision trees certainly did a better job of differentiating between the two classes. In comparison to the first tree, the second one could potentially overfit the training set.
Next, you saw the decision boundary of an SVM model.
The decision boundary of the SVM model was not linear because it used a radial kernel as shown below:

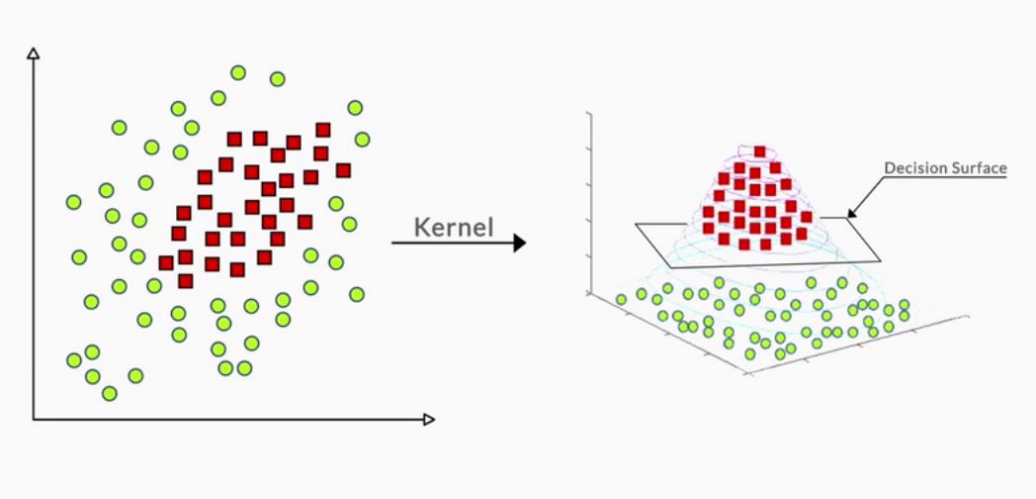

Figure 6 – Transformation using a kernel

The decision boundary of the SVM is shown below.

Figure 7 - SVM decision boundary

The SVM differentiates the two classes almost perfectly, using a hyperplane that looks like a circle when picturing it in a two-dimensional space. In this business problem, an SVM does a much better job than a decision tree or logistic regression.

Pros and Cons of Different Machine Learning Models

Let's look at the advantages and disadvantages of each of the algorithms covered so far.

Advantages

1. **Logistic regression**
    1. It offers convenient **probability scores** for outputs.
    2. It can be efficiently implemented across **different tools**.

3. The issue of **multicollinearity** can be countered with **regularisation**.
4. It has a **widespread industry use**.

2. **Decision trees**
    1. Intuitive decision rules make it **easier to interpret** the data.
    2. Trees handle **nonlinear features** well.
    3. The **variable interaction** is taken into account.

3. **Support vector machines**
    1. SVMs can handle **large feature spaces**.
    2. They can handle **nonlinear feature** interactions.
    3. They do not rely on the entire **dimensionality** of the data for the transformation.

Disadvantages

1. **Logistic regression**
    1. It does not perform well when the feature space is too large.
    2. It does not perform well when there are a lot of categorical variables in the data.
    3. The nonlinear features must be transformed into linear features to efficiently use them for a logistic model.
    4. It relies on the entire data; so if there is even a small change in the data, the logistic model can change significantly.

2. **Decision trees**
    1. Trees are highly biased towards the training set and **overfit** it quite often.
    2. There is **no probabilistic output** score as the output.

3. **Support vector machines**
    1. SVMs are not efficient, in terms of **computational cost**, when the number of observations is large.
    2. It is tricky and time-consuming to find the appropriate **kernel** for a given data set.

End-to-End Modelling - I

You could get overwhelmed by the choice of algorithms available for classification. But let's now look **at how you should go about modelling** data. To summarise —

1. Start with **logistic regression**. Using a logistic regression model serves two purposes:
    a. It acts as a **baseline** (benchmark) model.
    b. It gives you an idea about the **important variables**.

2. Then, opt for **decision trees** and compare their performance with the logistic regression model. If there is no significant improvement in their performance, then use the important variables drawn from the logistic regression model.

3. Finally, if you still do not meet the performance requirements, use **support vector machines**. But, keep in mind the **time and resource constraints**, because it takes time to find an appropriate kernel for SVM. Also, they are computationally expensive.

CART and CHAID Trees

So far, you studied a specific type of tree: **CART (Classification and Regression Tree)**. There is one more tree that is used widely. It is called **CHAID (Chi-square Automatic Interaction Detection)**. Both of these trees have different applications.

CART is used to create binary trees, i.e. trees that can have a maximum of two possible children for a node. It is best suited for prediction (supervised learning) tasks. But sometimes, CART is not appropriate to visualise the important features in a data set because binary trees tend to be much **deeper** and more **complex**.

CHAID trees are used to create non-binary trees, i.e. trees that can have more than two children for a node. This feature makes the trees wider rather than deeper; this, in turn, makes it easier to look at them and understand the important drivers (features) in a business problem. The process of finding out important features is also referred to as **driver analysis**.

To put CART and CHAID in the form of an analogy, suppose you are working with the Indian cricket team, and you want to **predict** whether the team will win a particular tournament or not. In this case, **CART** would be preferable because it is more suitable for prediction tasks. Whereas, if you want to look at the **factors** that are going to influence the win/loss of the team, then a **CHAID** tree would be more preferable.

Choosing between Trees and Random Forests

You learnt about decision trees and random forests. Let's now learn **how to choose between the two,** by looking at the disadvantages of decision trees and the advantages that random forests have over them.

Disadvantages of decision trees

1. Trees have a tendency to **overfit** the training data.
2. Splitting with **multiple linear decision boundaries** is not always efficient.
3. It is not possible to **predict beyond the range** of the response variable in the training data in a regression problem.

Suppose you want to predict house prices using a decision tree, and the price range of the house (response variable) is $5,000 to $35,000. While predicting, the output of the decision tree will always be within this range.

Advantages of random forests

1. There is no need to **prune** the trees of a forest.
2. The **OOB error** can be calculated from the training data itself, which gives a good estimate of model performance on unseen data.
3. It is hard for a random forest to **overfit** the training data.
4. A random forest is not affected by **outliers** too much because of the aggregation strategy.

Although random forests overcome most of the issues that trees face, forests also have some limitations. Some of the limitations include the following:

1. Given their origin in decision trees, random forests have a similar problem of **not predicting beyond the range of the response variable** in the training set.
2. The **extreme values are often not predicted** because of the aggregation strategy.

End-to-End Modelling - II

So far, you learnt about multiple machine learning models. They include —

1. Logistic regression
2. Decision trees
3. Support vector machines
4. Types of decision trees
5. Random forests

Sometimes, you may get overwhelmed and confused by the choice of models. In such a scenario, how do you go about modelling the data? Which model should you choose? Here's the answer. In general, you should start with a **logistic regression** model. Then, you should build a **decision tree** model. While building a decision tree, you should choose the appropriate method: **CART** for predicting and **CHAID** for a driver analysis. If you are not satisfied with the model performance mentioned so far and you have sufficient time and resources in hand, then you should go ahead and build more complex models such as **random forests** and **support vector machines**.