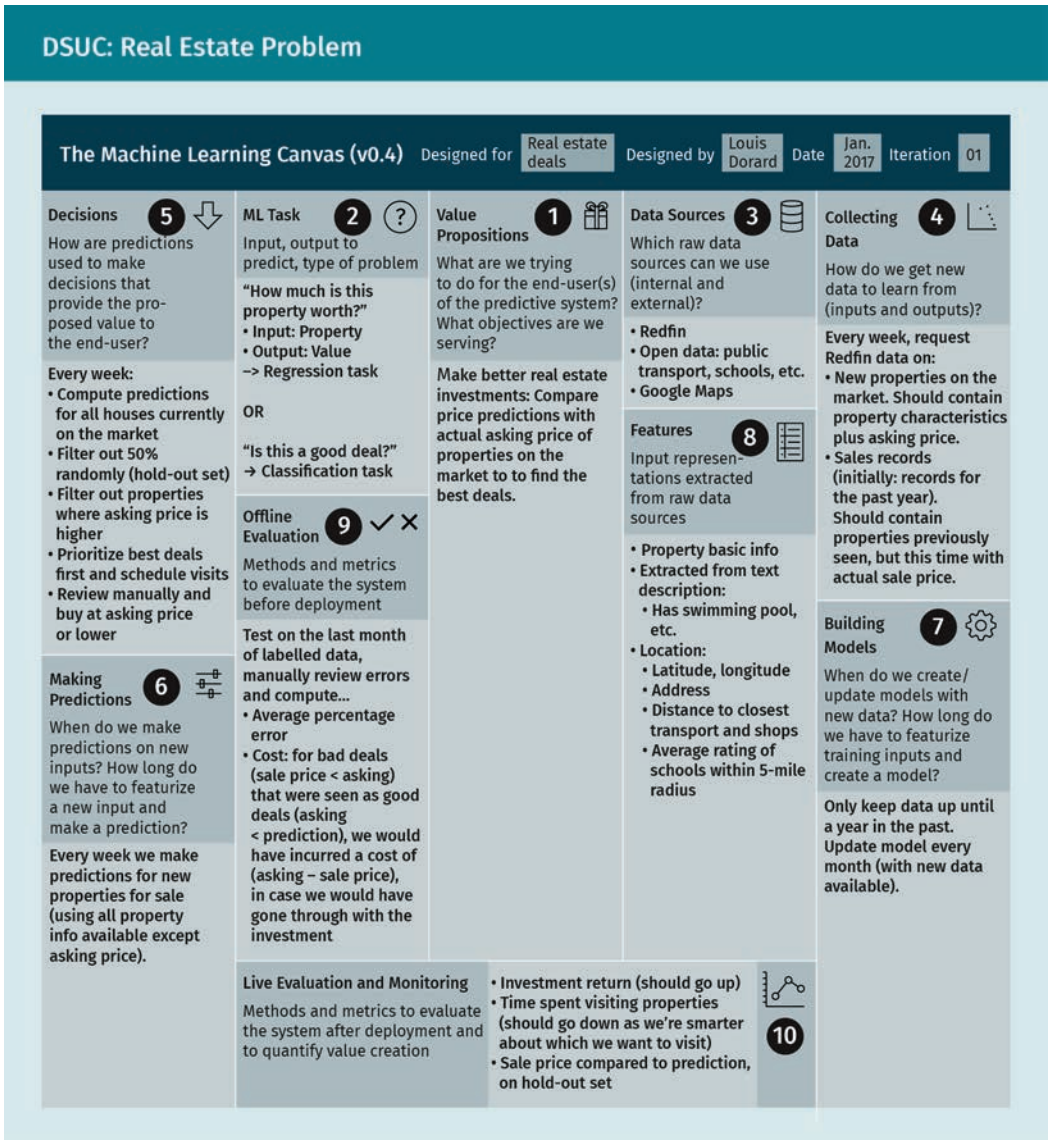


3.2 Performance Evaluation

There are two approaches that can be used to properly evaluate the success of a DSUC model and whether or not it accomplishes its business-oriented objectives. The first approach is to evaluate the model by comparing its output through a list of well-established numerical metrics. The second approach is to evaluate the ways that the model influenced the business by helping it to improve and achieve its goals.

Model-Centric Evaluation: Performance Metrics

The output of the developed prediction model is either a class or category (classification model), or a discrete number or probability (regression model). We will discuss the metrics that are routinely implemented to evaluate the performance of each type of these models.

Evaluation metrics for a classification model

For a DSUC designed with only two possible outputs {"yes", "no"}, the decision of the output is dependent on a threshold assigned to the model. When the model is applied to a data record, there are only four possible outcomes. These are true positive, true negative, false positive and false negative.

- True positive (TP): In the case of a correct prediction, the classifier produces a label "yes" for the data record.
- True negative (TN): In the case of a correct prediction, the classifier produces a label "no" for the data record.
- False positive (FP): In the case of an incorrect prediction, the classifier produces a label "yes" for the data record.
- False negative (FN): In the case of an incorrect prediction, the classifier produces a label "no" for the data record.

These four possible results are usually presented in a matrix form called the confusion matrix, as shown below.

The Confusion Matrix

		Model Output	
		YES	NO
Desired Output	YES	Number of TPs	Number of FNs
	NO	Number of FPs	Number of TNs

For the four possible outputs, there are three performance metrics to measure the model quality. These are precision, accuracy, and recall, as explained in the following equations.

$$\text{Precision} = \frac{\text{number of TP}}{\text{number of TP} + \text{number of FP}}$$

$$\text{Accuracy} = \frac{\text{number of TP} + \text{number of TN}}{\text{number of TP} + \text{number of TN} + \text{number of FP} + \text{number of FN}}$$

$$\text{Recall} = \frac{\text{number of TP}}{\text{number of TP+number of FN}}$$

To distinguish between the two classes of the classification model {“yes”, “no”}, a threshold has to be applied. This cutoff value could be set to a certain percentage that is decided upon during the analysis, and any output value that exceeds this cutoff value will be considered a “yes,” while all lower outputs will be considered a “no.” Therefore, the model performance is dependent on the cutoff value which affects the number of true positives, true negatives, false positives, and false negatives accordingly.

The receiver operator characteristic (ROC) curve shows how altering the cutoff value could change the true positive and false positive rates. An ideal model would be able to complete the classification operation with 100 percent accuracy, meaning that it could produce a true positive value of 100 percent and a false positive value of zero percent. Since no model in reality can be that accurate, the ROC curve helps to find a more realistic threshold value at which true positive is at its highest rate and false positive is at its lowest rate. The following steps should be followed to create a ROC curve:

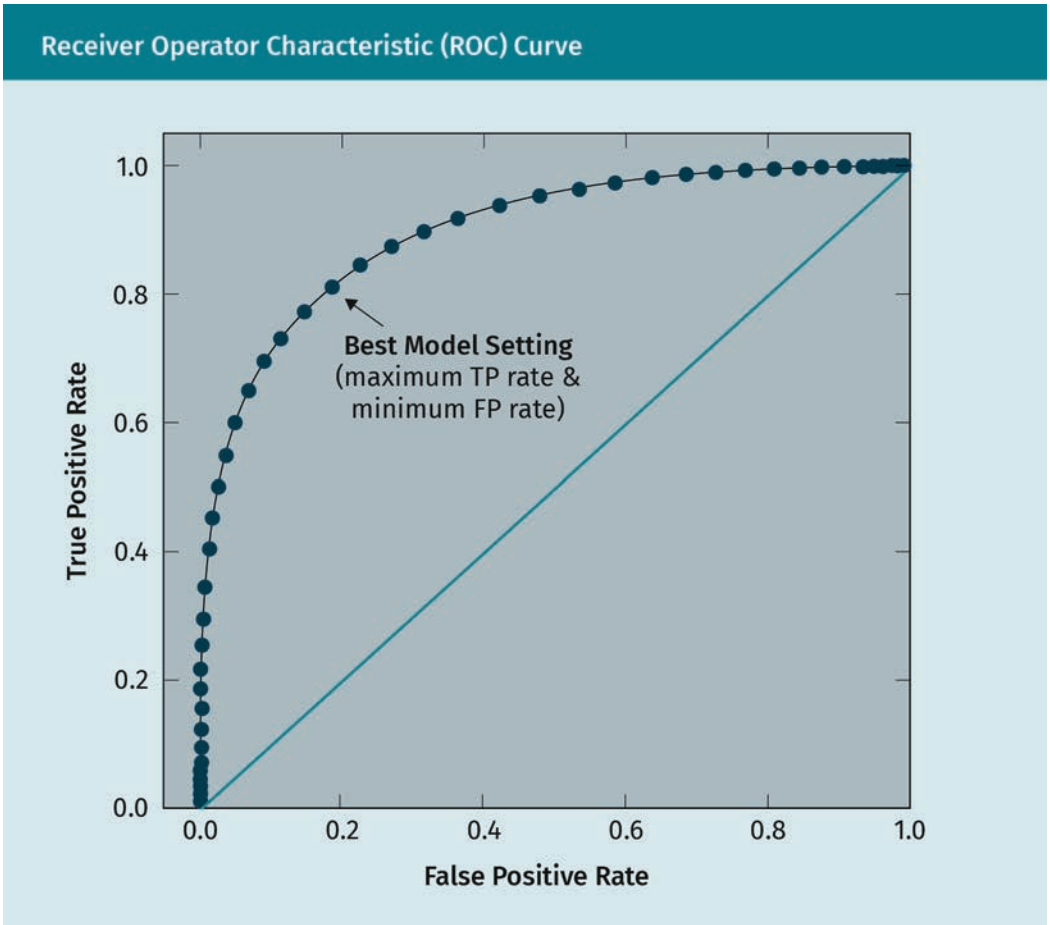
1. A cut off value has to be chosen ranging from 0 to 100.
2. The model is applied to a test set, and the numbers of TP, TN, FP and FN are recorded.
3. Calculate:

$$\text{False Positive Rate} = \frac{\text{number of FP}}{\text{number of FP+number of TN}}$$

and

$$\text{True Positive Rate} = \frac{\text{number of TP}}{\text{number of TP+number of FN}}$$

4. Every point on the ROC curve has coordinates of (False Positive Rate, True Positive Rate).
5. Another cutoff frequency is chosen, and steps 2 and 4 are repeated, resulting in the ROC shown below.



Evaluation metrics for a regression model

The objective is to measure how close a regression model's output (y) is to the desired output (d). There are standard metrics that evaluate the accuracy and performance of the model which are root mean square error, mean absolute error, absolute error, mean absolute error, relative error, and square error, as given in the following equations.

$$\text{Absolute error } (\varepsilon) = |d - y|$$

$$\text{Relative error } (\varepsilon^*) = \left|\frac{d-y}{d}\right| \cdot 100\,\%$$

$$\text{Mean absolute percentage error (MAPE)} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{d_i - y_i}{d_i}\right| \cdot 100\,\%$$

$$\text{Square error } (\varepsilon^2) = (d-y)^2$$

$$\text{Mean square error (MSE)} = \frac{1}{n}\sum_{i=1}^{n}(d_i - y_i)^2$$

$$\text{Mean absolute error (MAE)} = \frac{1}{n}\sum_{i=1}^{n}|d_i - y_i|$$

$$\text{Root mean square error (RMSE)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - y_i)^2}$$

3.3 Data-Driven Operational Decisions

A crucial aspect in the praxis of data science lies in the operationalization of the insights derived from the employed analytical models. To this end, it is of vital importance that analytics results are communicated and made available in such a way that they are useful for the relevant decision makers inside an organization. Moreover, it is usually helpful to explain the rationale behind the modeling approach so that the end-user can make an informed interpretation of model results.

The end user decides how to line up the model's output in order to fit in the business goals and objectives. For example, in fraud detection, the user has the ability to decide the range of percentages at which a suspicious transaction or behaviour is considered to be a true fraud. In this case, the selected threshold will have a tradeoff between false negatives and false positives. These tradeoffs should be taken into consideration to maximize the effectiveness of the model. Using different values for the threshold enables the business managers to consider different scenarios.

In some cases, the end user may have to make a decision that directly impacts data records, as well as a decision about the value of the thresholds. For example, one feature that may exist in a dataset is product price. In some cases the price of the product may need to be modified. On such an occasion, the model should be capable of accommodating these changes and should be able to be re-trained.

The ultimate goal of a smart model is the automation of user's decisions. These decisions are often dependent on the model's ability of prediction. For example, a model could be designed to analyze hotel reviews and decide whether these reviews are fake or not. If the model's predictions are highly accurate, then a review can automatically be accepted or rejected without the need for any human intervention.

Business-Centric Evaluation: The Role of KPIs

After a model has been evaluated successfully with the aforementioned evaluation metrics, it is ready for deployment. At this stage, the model should be able to produce a trusted DSUC value for the associated business problem. The decision makers must then be confident that the DSUC is correctly implemented in a way that will help the company to meet their business goals. Quantification of the model's merit is achieved by defining so-called Key Performance Indicators (KPIs). These are measurements that express to what extent the business goals have been met or not. Most KPIs focus on increased efficiency, reduced costs, improved revenue, and enhanced customer satisfaction.