

***k*-fold Cross Validation Decision Trees**

Using Student Measures and Excel

I. Prepare the data

- Assign letter grades to grade points

II. Review Excel Fundamentals

- Relative and Absolute Cell References
- Range Names
- VLOOKUP
- RANDBETWEEN

III. Create *k*-fold Samples, $k = 5$

- Create 5 random samples from 292 records
- Create 5 Training/Validation Samples

Assigning Letter Grades to Grade Points

$$Retention_Risk = \begin{cases} Low \\ High \end{cases}$$

$$\begin{cases} Num_Semem_To \leq 3 \\ Num_Semem_To > 3 \end{cases}$$

$$Retention_Risk = f(AvgGPChg_To, \text{Base_GPA}, AvgProGpa_To, AvgDelta_AttPass_To, AvgDiffPts_To, Age, Gender, Race) (2)$$

Grades and Grade Points

Letter Grade	Grade Points for each hour
A (Excellent)	4.0
A-	3.7
B+	3.3
B (Good)	3.0
B-	2.7
C+	2.3
C (Fair)	2.0
C-	1.7
D+	1.3
D	1.0
D- (Barely Passing)	0.7
F (Failure)	0.0

Excel Fundamentals

References: Relative vs. Absolute

Excel uses a technique called **relative cell address** when it copies a formula from one cell to another. When Excel copies or moves a formula, it may adjust the column letter(s) for each column. Formulas copied to different rows will also have the row addresses adjusted properly.

Using the relative cell address in copied formulas is appropriate most of the time and that is the default treatment. However, there are some instances when you do not want the cell address to be adjusted automatically. For example, when calculating percentages of a total, you want to divide each item by the same amount.

With references, we can:

- Identify cells or groups of cells
- Tell Excel which cells to look in to find values to be used in a formula

There are **three** types of cell references used in Excel:

- Relative
- Absolute
- Mixed

Because references are based on row and column headings in a worksheet, the dollar sign (\$) preceding the row or column reference indicates which absolute cell reference is desired. Let C denote a column letter, and R denote a row number.

CR Relative Reference. Both the row and column are adjusted when the expression is copied. This reference tells Excel how to find another cell, starting from the cell containing the formula.

\$C\$R Absolute Reference. Neither row nor column is adjusted when the expression is copied. Gives the exact location of the cell.

\$CR Mixed Reference. The column is not adjusted when the expression is copied, BUT the row is adjusted.

C\$R Mixed Reference. When the expression is copied, the row is NOT adjusted BUT the column is adjusted.

Besides typing in the references, you can select the reference in the formula bar and press the **F4** function key to toggle through the reference types (R, A, M_r, M_c)

Naming Ranges

Learn more about using names

A name is a meaningful shorthand that makes it easier to understand the purpose of a **cell reference**, **constant**, **formula**, or **table**, each of which may be difficult to comprehend at first glance. The following information shows common examples of names and how they can improve clarity and understanding.

Example Type	Example with no name	Example with a name
Reference	=SUM(C20:C30)	=SUM(FirstQuarterSales)
Constant	=PRODUCT(A5,8.3)	=PRODUCT(Price,WASalesTax)
Formula	=SUM(VLOOKUP(A1,B1:F20,5,FALSE),—G5)	=SUM(Inventory_Level,—Order_Amt)
Table	C4:G36	=TopSales06

Types of names

There are several types of names that you can create and use.

Defined name A name that represents a cell, range of cells, formula, or constant value. You can create your own defined name, and Microsoft Office Excel sometimes creates a defined name for you, such as when you set a print area.

Table name A name for an Excel table, which is a collection of data about a particular subject that is stored in records (rows) and fields (columns). Excel creates a default Excel table name of Table1, Table2, and so on, each time that you insert an Excel table, but you can change the name to make it more meaningful. For more information on Excel tables, see [Using structured references with Excel tables](#).

The scope of a name

All names have a scope, either to a specific worksheet (also called the local worksheet level) or to the entire workbook (also called the global workbook level). The scope of a name is the location within which the name is recognized without qualification. For example:

If you have defined a name, such as Budget_FY08, and its scope is Sheet1, then that name, if not qualified, is only recognized in Sheet1, but not in Sheet2 or Sheet3 without qualification.

To use a local worksheet name in another worksheet, you can qualify it by preceding it with the worksheet name, as the following example shows:

Sheet1!Budget_FY08

If you have defined a name, such as Sales_Dept_Goals, and its scope is the workbook, then that name is recognized for all worksheets in that workbook, but not for any other workbook.

A name must always be unique within its scope. Excel prevents you from defining a name that is not unique within its scope. However you can use the same name in different scopes. For example, you can define a name, such as GrossProfit, scoped to Sheet1, Sheet2, and Sheet3 in the same workbook. Although each name is the same, each name is unique within its scope. You might do this to ensure that a formula that uses the name, GrossProfit, is always referencing the same cells at the local worksheet level.

You can even define the same name, GrossProfit, for the global workbook level, but again the scope is unique. In this case, however, there can be a name conflict. To resolve this conflict, by default Excel uses the name that is defined for the worksheet, because the local worksheet level takes precedence over the global workbook level. If you want to override the precedence and you want to use the workbook name, you can disambiguate the name by prefixing the workbook name as the following example shows:

WorkbookFile!GrossProfit

You can override the local worksheet level for all worksheets in the workbook, with the exception of the first worksheet, which always uses the local name if there is a name conflict, and cannot be overridden.

Creating and entering names

You create a name by using the:

Name box on the formula bar This is best used for creating a workbook level name for a selected range.

Create a name from selection You can conveniently create names from existing row and column labels by using a selection of cells in the worksheet.

New Name dialog box This is best used for when you want more flexibility in creating names, such as specifying a local worksheet level scope or creating a name comment.

NOTE By default, names use **absolute cell references**.

You can enter a name by:

Typing Typing the name, for example, as an argument to a formula.

Using Formula AutoComplete Use the Formula AutoComplete drop-down list, where valid names are automatically listed for you.

Selecting from the Use in Formula command Select a defined name from a list available from the **Use in Formula** command in the **Defined Names** group on the **Formulas** tab.

Auditing names

You can also create a list of defined names in a workbook. Locate an area with two empty columns on the worksheet (the list will contain two columns, one for the name and one for a description of the name). Select a cell that will be the upper-left corner of the list. On the **Formulas** tab, in the **Defined Names** group, click **Use in Formula**, click **Paste**, and then in the **Paste Names** dialog box, click **Paste List**.

[↕ Top of Page](#)

Learn about syntax rules for names

The following is a list of syntax rules that you need to be aware of when you create and edit names.

Valid characters The first character of a name must be a letter, an underscore character (**_**), or a backslash (****). Remaining characters in the name can be letters, numbers, periods, and underscore characters.

NOTE You cannot use the letters "C", "c", "R", or "r" as a defined name, because both of these letters are used as a shorthand for selecting a row or column for the currently selected cell when you enter them in a **Name** or **Go To** text box.

Cell references disallowed Names cannot be the same as a cell reference, such as Z\$100 or R1C1.

Spaces are not valid Spaces are not allowed. Use the underscore character (_) and period (.) as word separators, such as, Sales_Tax or First.Quarter.

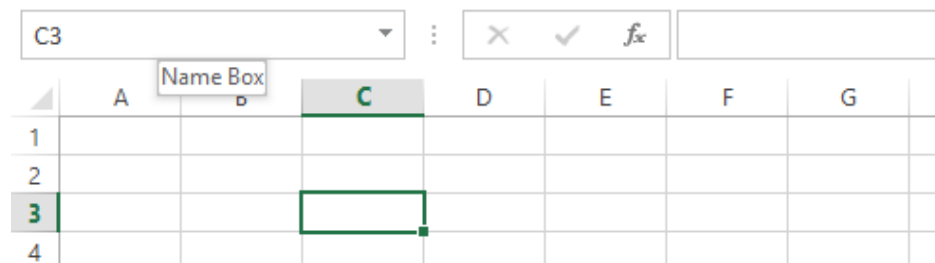
Name length A name can contain up to 255 characters.

Case sensitivity Names can contain uppercase and lowercase letters. Excel does not distinguish between uppercase and lowercase characters in names. For example, if you created the name Sales and then create another name called SALES in the same workbook, Excel prompts you to choose a unique name.

[↑ Top of Page](#)

Create a name for a cell or cell range on a worksheet

1. Select the cell, range of cells, or **nonadjacent selections** that you want to name.
2. Click the **Name** box to the left of the **formula bar**.



Name box

3. Type the name that you want to use to refer to your selection. Names can be up to 255 characters in length.
4. Press ENTER.

NOTE You cannot name a cell while you are changing the contents of the cell.

[↑ Top of Page](#)

Create a name by using a selection of cells in the worksheet

You can convert existing row and column labels to names.



1. Select the range that you want to name, including the row or column labels.
2. On the **Formulas** tab, in the **Defined Names** group, click **Create from Selection**.
3. In the **Create names from Selection** dialog box, designate the location that contains the labels by selecting the **Top row**, **Left column**, **Bottom row**, or **Right column** check box.

NOTE A name created by using this procedure refers only to the cells that contain values and does not include the existing row and column labels.

[Top of Page](#)

Create a name by using the New Name dialog box

1. On the **Formulas** tab, in the **Defined Names** group, click **Define Name**.
2. In the **New Name** dialog box, in the **Name** box, type the name that you want to use for your reference. Names can be up to 255 characters in length.
3. To specify the scope of the name, in the **Scope** drop-down list box, select **Workbook**, or the name of a worksheet in the workbook.
4. Optionally, enter a descriptive comment up to 255 characters.
5. In the **Refers to** box, do one of the following:

Cell reference The current selection is entered by default. To enter other cell references as an argument, click **Collapse Dialog**  (which temporarily hides the dialog box), select the cells on the worksheet, and then click **Expand Dialog** .

Constant Type an = (equal sign), followed by the constant value.

Formula Type an = (equal sign) followed by the formula.

6. To finish and return to the worksheet, click **OK**.

TIP To make the **New Name** dialog box wider or longer, click and drag the grip handle at the bottom.

[Top of Page](#)



Manage names by using the Name Manager dialog box

Use the **Name Manager** dialog box to work with all of the defined names and table names in the workbook. For example, you may want to find names with errors, confirm the value and reference of a name, view or edit descriptive comments, or determine the scope. You can also sort and filter the list of names, and easily add, change, or delete names from one location.

To open the **Name Manager** dialog box, on the **Formulas** tab, in the **Defined Names** group, click **Name Manager**.

View names

The **Name Manager** dialog box displays the following information about each name in a list box:

This Column:	Displays:
Icon and Name	One of the following: A defined name, which is indicated by a defined name icon.  A table name, which is indicated by a table name icon. 
Value	The current value of the name, such as the results of a formula, a string constant, a cell range, an error, an array of values, or a placeholder if the formula cannot be evaluated. The following are representative examples: "this is my string constant" 3.1459 {2003;12,2002;23,;2001,18} #REF! {...}
Refers To	The current reference for the name. The following are representative examples: =Sheet1!\$A\$3 =8.3 =HR!\$A\$1:\$Z\$345 =SUM(Sheet1!A1,Sheet2!B2)

Scope A worksheet name, if the scope is the local worksheet level.
"Workbook", if the scope is the global worksheet level.

Comment Additional information about the name up to 255 characters. The following are representative examples:

This value will expire on May 2, 2007.

Don't delete! Critical name!

Based on the ISO certification exam numbers.

NOTE If you save the workbook to Microsoft Office SharePoint Server 2007 Excel Services, and you specify one or more parameters, the comment is used as a tooltip in the **Parameters** toolpane.

NOTES

You cannot use the **Name Manager** dialog box while you are changing the contents of the cell.

The **Name Manager** dialog box does not display names defined in Visual Basic for Applications (VBA), or hidden names (the **Visible** property of the name is set to "False").

VLOOKUP Function

This useful functions allows the user to match a value in a table, and will "look up" a corresponding value from that table. That is, VLOOKUP compares a search value with the first column in a list and returns an associated value in the same row.

The syntax for the VLOOKUP function is:

=VLOOKUP(*lookup_value*, *table_array*, *col_index_num*, *range_lookup*)

where

lookup_value specifies the search value you want to compare with the **first** column in the list.

table_array is a cell range or range name for the lookup table, the first column of which contains the lookup values.

col_index_num gives the number of the columns in the array where the value or label is to be retrieved.

range_lookup argument specifies how to compare the search value with the first column. Entering FALSE tells Excel to find an exact match for the search value. TRUE tells Excel to find an approximate match.

Note: The VLOOKUP function requires the lookup column to be the first column of your list. If you want to do a range lookup and the lookup column is not the first column of the list, then you must use the vector form of the LOOKUP function. For example, the syntax

=Lookup(J4, 'Grading Scale'!A3:A14, 'Grading Scale'!B3:B14)

compares search values to the Range column values and returns the associated alue in the Grade column, regardless of the order of the columns in the list.

Example:

It is rather easy to use the VLOOKUP function to assign letter grades to numerical scores. Suppose my grading scale is as follows:

Range	Grade
< 55	F
55 - 62	D-
62 - 68	D
68 - 70	D+
70 - 72	C-
72 - 78	C
78 - 80	C+
80 - 82	B-
82 - 88	B
88 - 90	B+
90 - 92	A-

≥ 92	A
-----------	---

Before we can implement the VLOOKUP function, we need to set up the table or an array defining the grading scale.

In searching for the perfect match, the list containing the values that you want to search much be in ascending order. This is because Excel finds the first value greater than your search value and then backs up one position and returns the associated value.

In Excel, the Grading Scale table would be as follows:

Range	Grade
0	F
55	D-
62	D
68	D+
70	C-
72	C
78	C+
80	B-
82	B
88	B+
90	A-
92	A

Note: The ranges are defined as follows:

F: $0 \leq \text{score} < 55$

D-: $55 \leq \text{score} < 62$

.

.

.

A-: $90 \leq \text{score} < 92$

A: $\text{score} \geq 92$

If your lookup table is set up so that the lookup values are in the first row of a table rather than the first column, you can use the HLOOKUP function instead of the VLOOKUP function. The HLOOKUP function works identically to the VLOOKUP function, except that the table appears on its side, horizontally.

RANDBETWEEN function

Returns a random integer number between the numbers you specify. A new random integer number is returned every time the worksheet is calculated.

Syntax

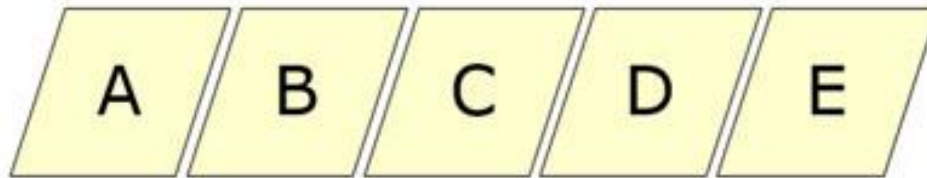
RANDBETWEEN(Bottom, Top)

The RANDBETWEEN function syntax has the following arguments:

- **Bottom** Required. The smallest integer RANDBETWEEN will return.
- **Top** Required. The largest integer RANDBETWEEN will return.

Note: When a worksheet is recalculated by entering a formula or data in a different cell, or by manually recalculating (press F9), a new random number is generated for any formula that uses the RANDBETWEEN function.

Create k -fold Training/Validation Samples, $k = 5$



	<u>Train</u>	<u>Validate</u>
1)	BCDE	A
2)	ACDE	B
3)	ABDE	C
4)	ABCE	D
5)	ABCD	E

Use a **programmatic approach for large samples**. For example, create a procedure using VBA that requires two arguments: (1) a range, and, (2) a value for k . The range should be a reference to the cells for the entire collection of data to use. The procedure should follow the general outline provided by the steps below.

STEPS:

- Determine the sample size, n , for samples A, B, C, D, and E.
 - We have 292 total student records, and $k = 5$, so, $n = 292/5 = 58$ (round down)
- Insert a column on the left edge of the table of all records. Label the column 'ID'.
- Enter values in the column for the 'ID', from 1 to 292.
- Insert a row at the top of the records. Label the row 'Column #'.
- Enter the values in the row for 'Column #', the value should equal the column number in the table. Be sure to number the blank columns included in the table.
- Assign a range name to the table of records. Do not include the row containing the column # in the reference. I used the range-name 'RawData'.
- Create k -fold samples
 - On a separate spreadsheet, label the columns for n and the sample IDs for samples A, B, C, D, and E.
 - Number the values of n appropriately.

- c. Enter the following Excel formula for each sample ID where $n = 1$:
=RANDBETWEEN(1, 292).
 - d. Copy the formula across the remaining columns and down the next 57 rows corresponding to $n = 2$, through $n = 58$.
 - e. Select the entire range of data, copy it, and paste_special—values. This freezes the data so that will not change when the spreadsheet is recalculated.
8. Create a spreadsheet for each training data set and each validation data set. Add labels for observation #, the ID, and all other fields of data being retrieved from the table. The column number for each field should appear on the row above.
 9. Number observations consecutively.
 10. Copy-and-paste the IDs from the appropriate samples A, B, C, D, and E, for each training and validation sample, appropriately. The training sample includes $(k-1) * n$ IDs (= 232), and the validation sample contains n IDs.
 11. Enter the following Excel formula in the cell next to the ID for the first observation: =VLOOKUP(\$B3,RawData,C\$1). Here, the IDs are stored in column 'B', and the column numbers are stored in row 1.
 12. Copy the formula to the remaining columns in the table.
 13. Copy the row of formulas to the remaining rows in the table.
 14. Repeat steps 11 – 13 for the remaining training and validation samples.
 15. Save the workbook.