National College of Ireland

Configuration Manual

MSc Research Project
Data Analytics

Elmarie Fryer
Student ID: 20220278

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Elmarie Fryer
Student ID:	20220278
Programme:	MSc in Data Analytics TopUp Module **Year:** 1
Module:	Top-up Module
Lecturer:	Jorge Basilio
Submission Due Date:	23rd August 2021
Project Title:	Predicting the likelihood of the need to launch a RNLI rescue boat in Ireland based on the Weather and Bank holidays Configuration Manual
	…… **Page Count:**
Word Count:	…….……

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……

23rd August 2021

Date: ……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Elmarie Fryer
Student ID: 20220278

1 Project Data

1.1 Data Sources

RNLI Data: https://data-rnli.opendata.arcgis.com/datasets/rnli-returns-of-service/explore
Irish Weather Data: https://www.kaggle.com/conorrot/irish-weather-hourly-data
Bank Holiday Data:
https://www.citizensinformation.ie/en/employment/employment_rights_and_conditions/leave_and_holidays/public_holidays_in_ireland.html#

ETL for Irish Weather Data
Source: Kaggle Dataset
Import as SQL table with fields as type nvarchar(50)
Data for later years has a space instead of NULL values
Convert to integer and float values
Calculate year, month, day, etc.
Delete values outside the range of the study, i.e. <2008 and >2019

1.2 Transformations

Irish Weather Stations [1] full list of 2079 stations in use since 1829. Select 25 active stations using a text editor.

Rename stations to match dataset

name	KaggleDataSetName
DUNSANY (Grange)	DUNSANY
JOHNSTOWN CASTLE	JOHNSTOWNII
MARKREE CASTLE	MARKREE
NEWPORT (Furnace)	NEWPORT
SHERKIN ISLAND	SherkinIsland

1.3 SQL Scripts

Use SQL Scripts RNLI001, RNLI003, RNLI006, RNLI008, RNLI012 to upload data
Use SQL Scripts RNLI002 to RNLI005 to prepare weather data
Use SQL Scripts RNLI007 to transform and update RNLI data

[1] https://www.met.ie/climate/available-data/historical-data

Use SQL Scripts RNLI008 through RNLI015 to further transform, extract partial data and create tables to query.

Note: data tables may take some time to upload as they are quite large.
The final data table (v03 grouped for Ireland is included in the Rapidminer artefact)

2 Calculate Closest Weather Station

Step 8 – calculate euclidean distances
Step 9 – generate csv and load into sql

Calculations are included in the excel files in the artefact

3 Initial Analysis

Replace UNKNOWN or NIL with blanks. 65k values replaced

Run Categorisation and Correlation Matrix in Rapidminer.
Saved process included in artifact

3.1 Correlation

	AvgTemp	BankHolidayWeekend = no	BoatLaunched = no	GroupedCategory = LEISURE	GroupedCategory = no incidents	GroupedOutcome = False Alarm / Hoax / Not found	GroupedOutcome = no incidents	GroupedOutcome = Others assistance	GroupedOutcome = RNLI Assistance	GroupedType = Boat	GroupedType = no incidents	GroupedType = Other	GroupedType = Waterside	HourGroupString = afternoon	HourGroupString = evening	HourGroupString = morning	Humidity	RoSTypeGrouped = ALB	RoSTypeGrouped = ILB	RoSTypeGrouped = no incidents	Season = Autumn	Season = Spring	Season = Summer	VisibilityGoodvBad = Bad	WindSorW = yes	Windspeed = LightModerate
AvgTemp	1.00	0.05	-0.20	0.20	-0.20	0.03	-0.20	0.07	0.17	0.16	-0.20	0.04	0.06	0.15	-0.01	-0.05	-0.04	0.11	0.16	-0.20	-0.02	-0.17	0.47	-0.00	0.09	0.06
BankHolidayWeekend = no	0.05	1.00	-0.19	0.16	-0.19	0.07	-0.19	0.07	0.14	0.12	-0.19	0.08	0.09	0.03	-0.00	-0.01	-0.10	0.13	0.12	-0.19	0.08	-0.06	0.02	-0.01	0.01	-0.06
BoatLaunched = no	-0.20	-0.19	1.00	-0.90	1.00	-0.35	1.00	-0.36	-0.78	-0.68	1.00	-0.39	-0.46	-0.21	0.02	0.07	-0.01	-0.66	-0.69	1.00	0.03	0.04	-0.17	-0.12	-0.03	-0.19
GroupedCategory = LEISURE	0.20	0.16	-0.90	1.00	-0.90	0.34	-0.90	0.36	0.67	0.47	-0.90	0.43	0.51	0.20	-0.00	-0.08	0.00	0.51	0.69	-0.90	-0.05	-0.04	0.18	0.10	0.03	0.17
GroupedCategory = no incidents	-0.20	-0.19	1.00	-0.90	1.00	-0.35	1.00	-0.36	-0.78	-0.68	1.00	-0.39	-0.46	-0.21	0.02	0.07	-0.01	-0.66	-0.69	1.00	0.03	0.04	-0.17	-0.12	-0.03	-0.19
GroupedOutcome = False Alarm / Hoax / Not found	0.03	0.07	-0.35	0.34	-0.35	1.00	-0.35	-0.03	-0.06	0.10	-0.35	0.32	0.19	0.03	-0.00	0.03	0.03	0.18	0.29	-0.35	0.00	0.01	0.00	0.04	0.02	0.07
GroupedOutcome = no incidents	-0.20	-0.19	1.00	-0.90	1.00	-0.35	1.00	-0.36	-0.78	-0.68	1.00	-0.39	-0.46	-0.21	0.02	0.07	-0.01	-0.66	-0.69	1.00	0.03	0.04	-0.17	-0.12	-0.03	-0.19
GroupedOutcome = Others assistance	0.07	0.07	-0.36	0.36	-0.36	-0.03	-0.36	1.00	-0.06	0.14	-0.36	0.22	0.24	0.07	0.00	-0.03	0.01	0.20	0.29	-0.36	-0.02	-0.02	0.07	0.05	0.02	0.07
GroupedOutcome = RNLI Assistance	0.17	0.14	-0.78	0.67	-0.78	-0.06	-0.78	-0.06	1.00	0.66	-0.78	0.17	0.31	0.18	-0.02	-0.07	-0.00	0.57	0.48	-0.78	-0.03	-0.04	0.16	0.09	0.02	0.15
GroupedType = Boat	0.16	0.12	-0.68	0.47	-0.68	0.10	-0.68	0.14	0.66	1.00	-0.68	-0.06	-0.07	0.16	-0.03	-0.04	-0.00	0.49	0.42	-0.68	0.00	-0.04	0.14	0.07	0.01	0.14

	AvgTemp	BankHolidayWeekend = no	BoatLaunched = no	GroupedCategory = LEISURE	GroupedCategory = no incidents	GroupedOutcome = False Alarm / Hoax / Not found	GroupedOutcome = no incidents	GroupedOutcome = Others assistance	GroupedOutcome = RNLI Assistance	GroupedType = Boat	GroupedType = no incidents	GroupedType = Other	GroupedType = Waterside	HourGroupString = afternoon	HourGroupString = evening	HourGroupString = morning	Humidity	RoSTypeGrouped = ALB	RoSTypeGrouped = ILB	RoSTypeGrouped = no incidents	Season = Autumn	Season = Spring	Season = Summer	VisibilityGoodvBad = Bad	WindSorW = yes	Windspeed = LightModerate
GroupedType = no incidents	-0.20	-0.19	1.00	-0.90	1.00	-0.35	1.00	-0.36	-0.78	-0.68	1.00	-0.39	-0.46	-0.21	0.02	0.07	-0.01	-0.66	-0.69	1.00	0.03	0.04	-0.17	-0.12	-0.03	-0.19
GroupedType = Other	0.04	0.08	-0.39	0.43	-0.39	0.32	-0.39	0.22	0.17	-0.06	-0.39	1.00	-0.04	0.04	0.00	-0.00	0.03	0.27	0.26	-0.39	-0.02	-0.00	0.03	0.08	0.02	0.07
GroupedType = Waterside	0.06	0.09	-0.46	0.51	-0.46	0.19	-0.46	0.24	0.31	-0.07	-0.46	-0.04	1.00	0.09	-0.00	-0.04	0.01	0.28	0.34	-0.46	-0.03	-0.01	0.06	0.05	0.03	0.08
HourGroupString = afternoon	0.15	0.03	-0.21	0.20	-0.21	0.03	-0.21	0.07	0.18	0.16	-0.21	0.04	0.09	1.00	-0.39	-0.39	-0.16	0.11	0.17	-0.21	-0.01	0.00	0.05	-0.00	0.01	0.01
HourGroupString = evening	-0.01	-0.00	0.02	-0.00	0.02	-0.00	0.02	0.00	-0.02	-0.03	0.02	0.00	-0.00	-0.39	1.00	-0.30	0.04	-0.02	-0.01	0.02	-0.02	0.00	0.02	-0.00	-0.01	0.02
HourGroupString = morning	-0.05	-0.01	0.07	-0.08	0.07	0.03	0.07	-0.03	-0.07	-0.04	0.07	-0.00	-0.04	-0.39	-0.30	1.00	0.06	-0.04	-0.05	0.07	0.01	0.01	-0.04	0.01	-0.00	-0.01
Humidity	-0.04	-0.10	-0.01	0.00	-0.01	0.03	-0.01	0.01	-0.00	-0.00	-0.01	0.03	0.01	-0.16	0.04	0.06	1.00	0.03	-0.01	-0.01	0.06	-0.10	-0.07	-0.05	0.06	-0.05
RoSTypeGrouped = ALB	0.11	0.13	-0.66	0.51	-0.66	0.18	-0.66	0.20	0.57	0.49	-0.66	0.27	0.28	0.11	-0.02	-0.04	0.03	1.00	-0.10	-0.66	-0.01	-0.03	0.09	0.07	0.03	0.11
RoSTypeGrouped = ILB	0.16	0.12	-0.69	0.69	-0.69	0.29	-0.69	0.29	0.48	0.42	-0.69	0.26	0.34	0.17	-0.01	-0.05	-0.01	-0.10	1.00	-0.69	-0.04	-0.03	0.14	0.08	0.02	0.15
RoSTypeGrouped = no incidents	-0.20	-0.19	1.00	-0.90	1.00	-0.35	1.00	-0.36	-0.78	-0.68	1.00	-0.39	-0.46	-0.21	0.02	0.07	-0.01	-0.66	-0.69	1.00	0.03	0.04	-0.17	-0.12	-0.03	-0.19
Season = Autumn	-0.02	0.08	0.03	-0.05	0.03	0.00	0.03	-0.02	-0.03	0.00	0.03	-0.02	-0.03	-0.01	-0.02	0.01	0.06	-0.01	-0.04	0.03	1.00	-0.36	-0.32	0.00	0.01	-0.01
Season = Spring	-0.17	-0.06	0.04	-0.04	0.04	0.01	0.04	-0.02	-0.04	-0.04	0.04	-0.00	-0.01	0.00	0.00	0.01	-0.10	-0.03	-0.03	0.04	-0.36	1.00	-0.37	-0.00	-0.07	0.00

	AvgTemp	BankHolidayWeekend = no	BoatLaunched = no	GroupedCategory = LEISURE	GroupedCategory = no incidents	GroupedOutcome = False Alarm / Hoax / Not found	GroupedOutcome = no incidents	GroupedOutcome = Others assistance	GroupedOutcome = RNLI Assistance	GroupedType = Boat	GroupedType = no incidents	GroupedType = Other	GroupedType = Waterside	HourGroupString = afternoon	HourGroupString = evening	HourGroupString = morning	Humidity	RoSTypeGrouped = ALB	RoSTypeGrouped = ILB	RoSTypeGrouped = no incidents	Season = Autumn	Season = Spring	Season = Summer	VisibilityGoodvBad = Bad	WindSorW = yes	Windspeed = LightModerate
Season = Summer	0.47	0.02	-0.17	0.18	-0.17	0.00	-0.17	0.07	0.16	0.14	-0.17	0.03	0.06	0.05	0.02	-0.04	-0.07	0.09	0.14	-0.17	-0.32	-0.37	1.00	0.01	0.05	0.11
VisibilityGoodvBad = Bad	-0.00	-0.01	-0.12	0.10	-0.12	0.04	-0.12	0.05	0.09	0.07	-0.12	0.08	0.05	-0.00	-0.00	0.01	-0.05	0.07	0.08	-0.12	0.00	-0.00	0.01	1.00	-0.02	-0.06
WindSorW = yes	0.09	0.01	-0.03	0.03	-0.03	0.02	-0.03	0.02	0.02	0.01	-0.03	0.02	0.03	0.01	-0.01	-0.00	0.06	0.03	0.02	-0.03	0.01	-0.07	0.05	-0.02	1.00	-0.07
Windspeed = LightModerate	0.06	-0.06	-0.19	0.17	-0.19	0.07	-0.19	0.07	0.15	0.14	-0.19	0.07	0.08	0.01	0.02	-0.01	-0.05	0.11	0.15	-0.19	-0.01	0.00	0.11	-0.06	-0.07	1.00

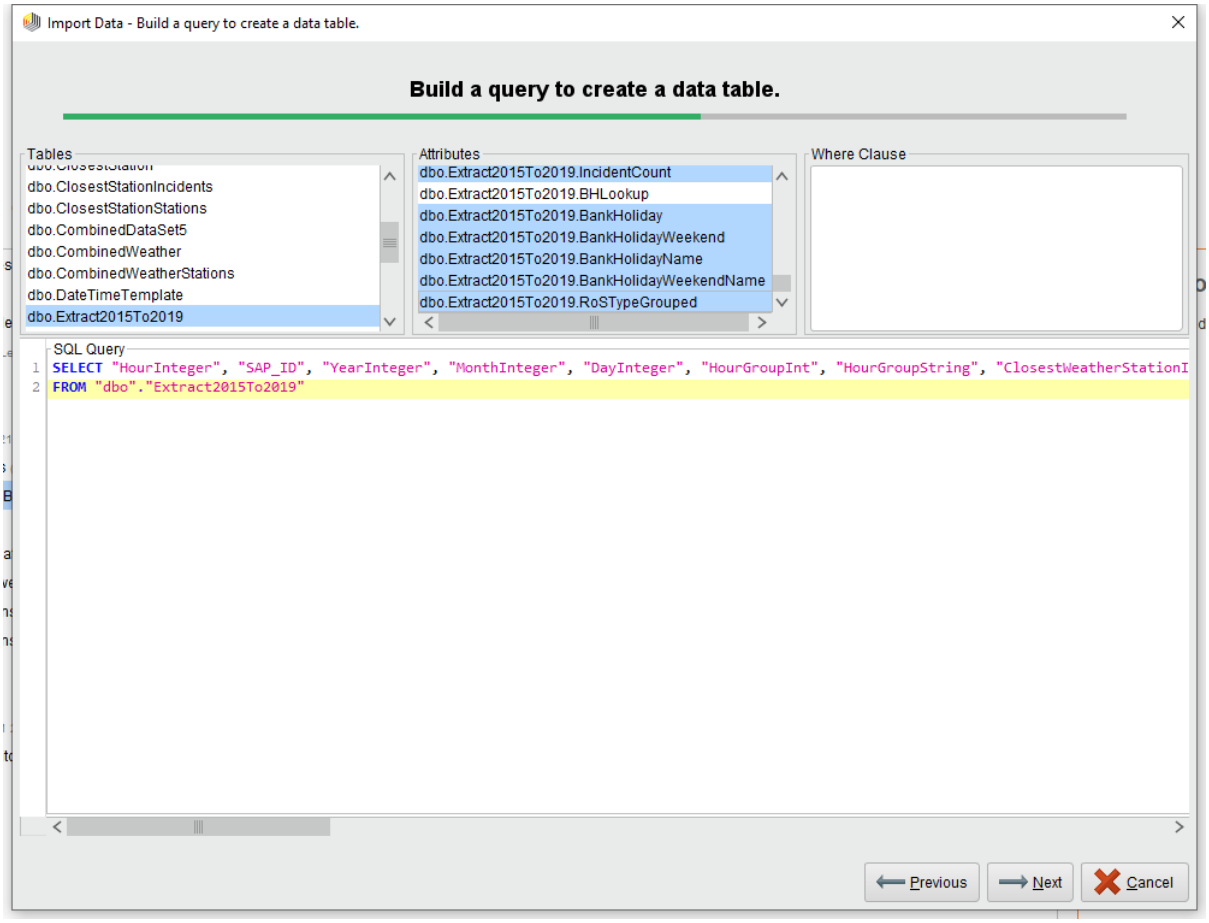
3.2 Categorisation

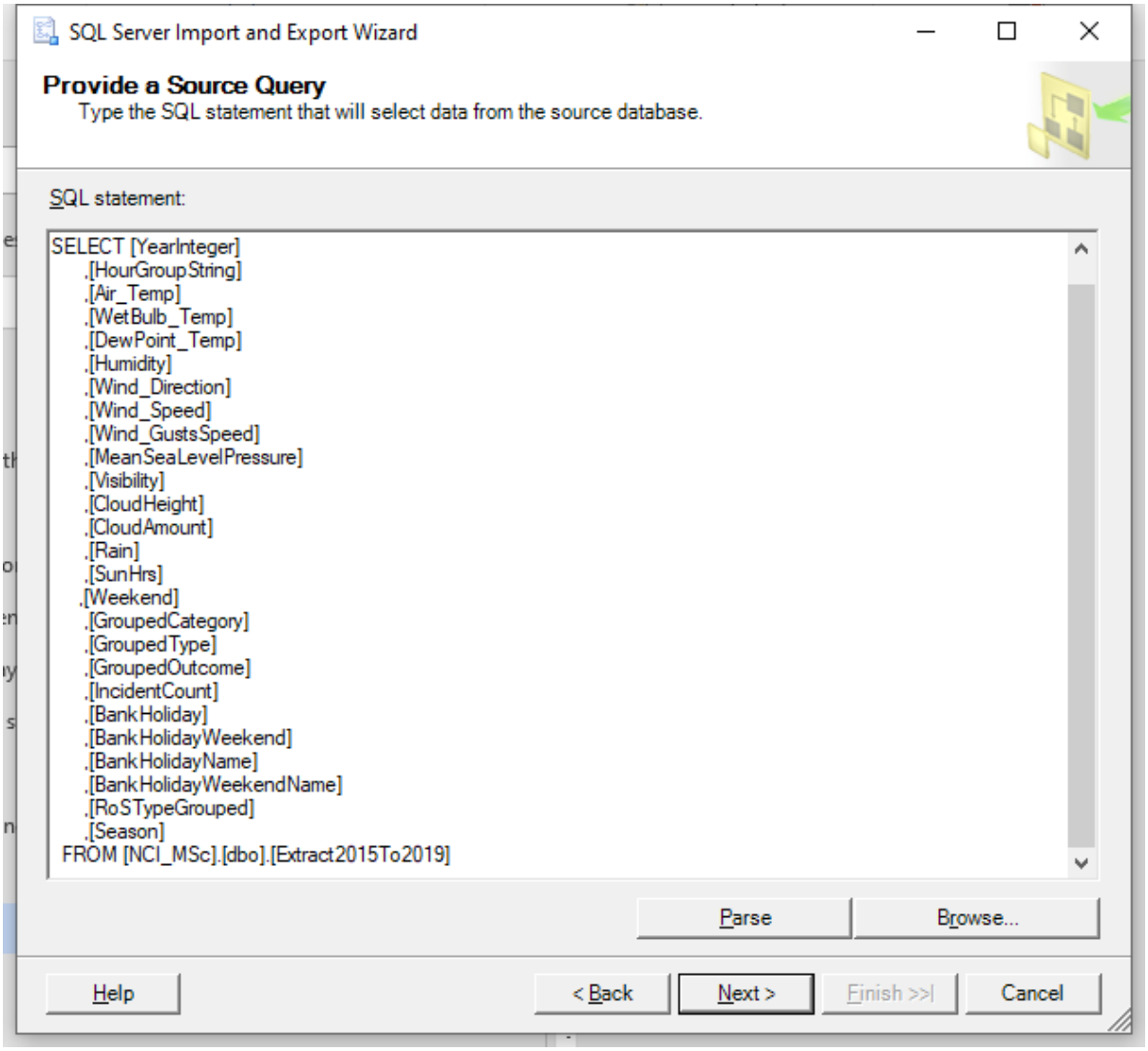
There is a high correlation between results and attributes of the result - no impact on prediction – leave them out.

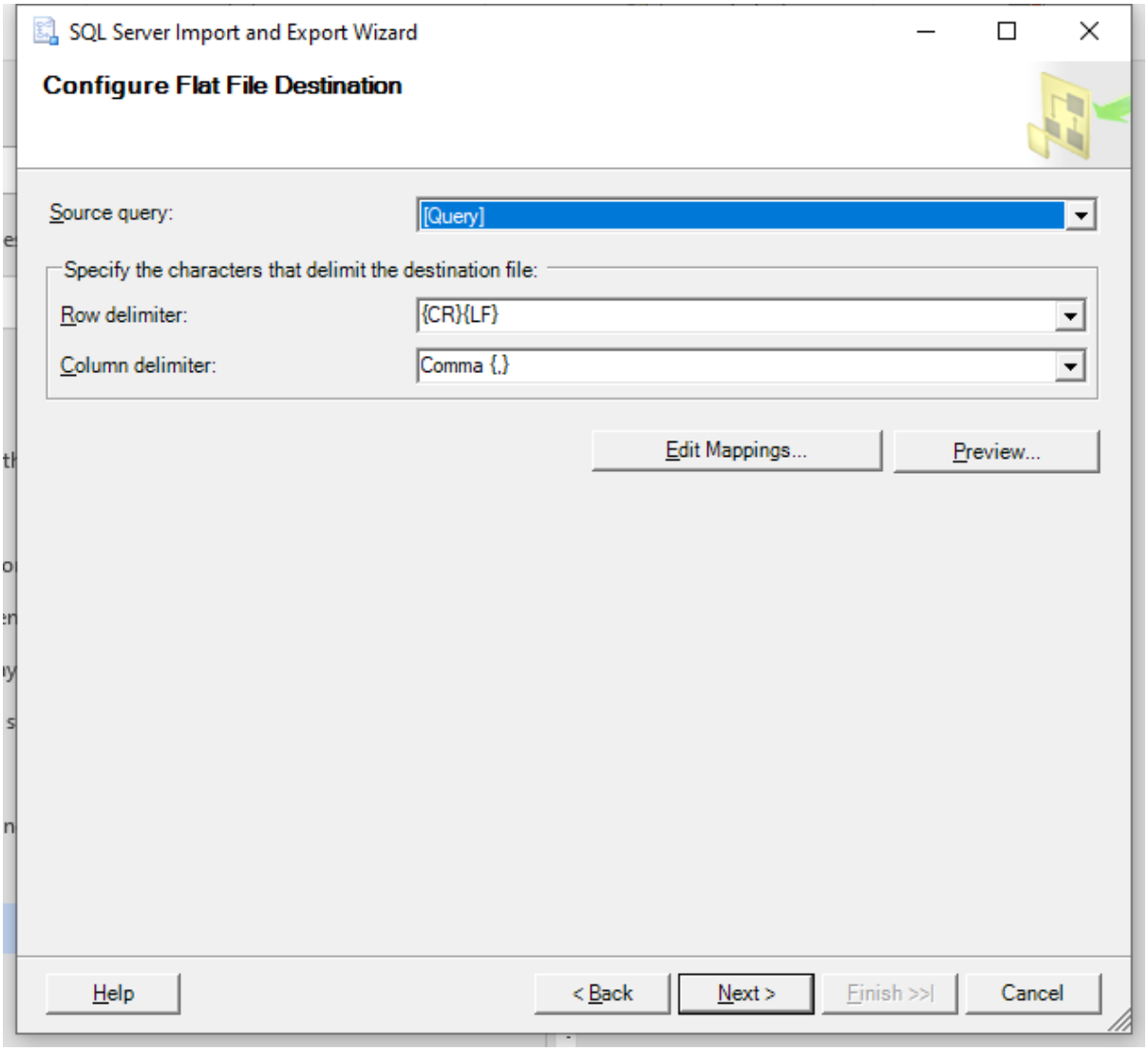
4 Rapidminer

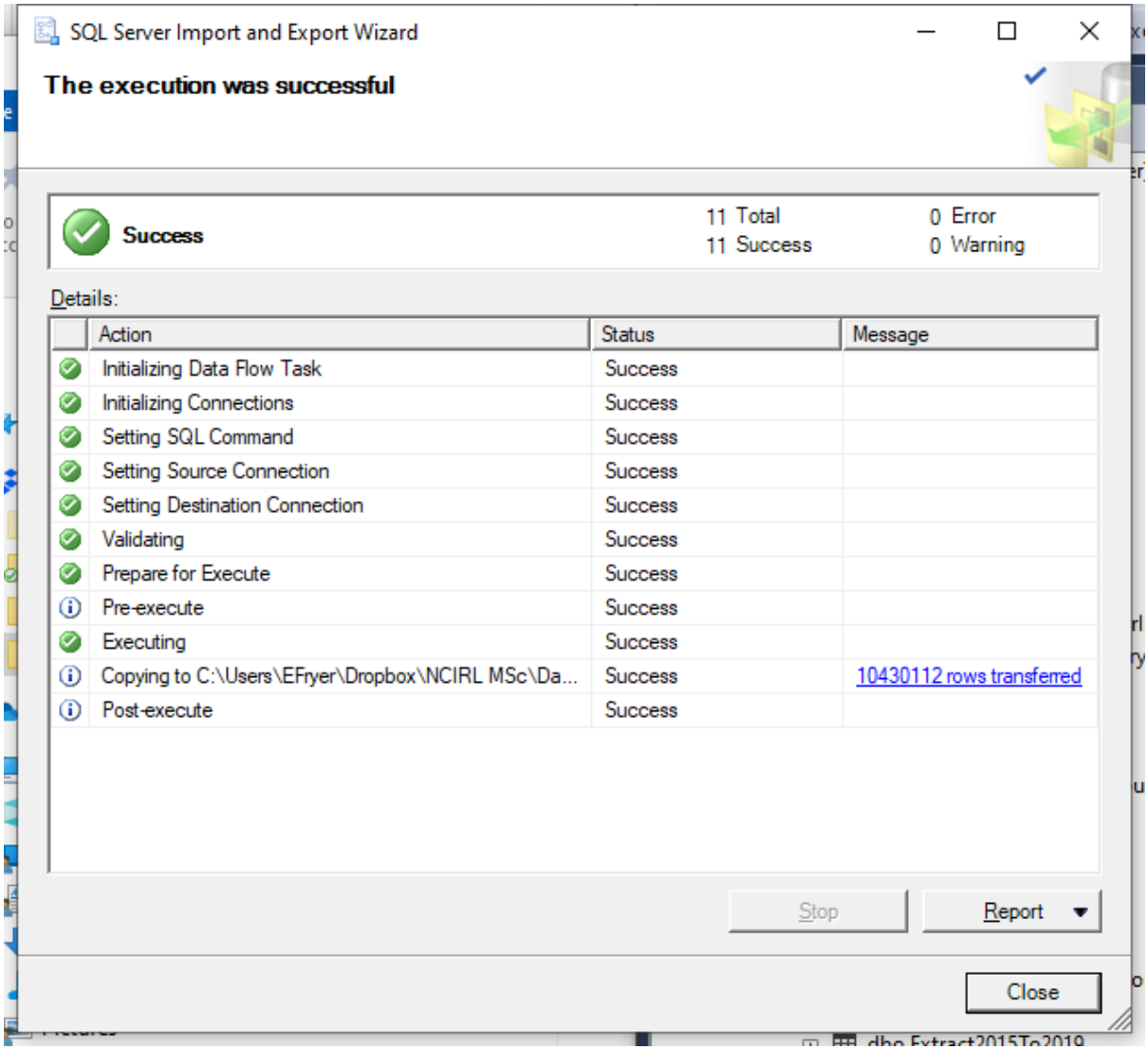
Import data from the sql server
Then use auto model to analyse the data and calculate the models.

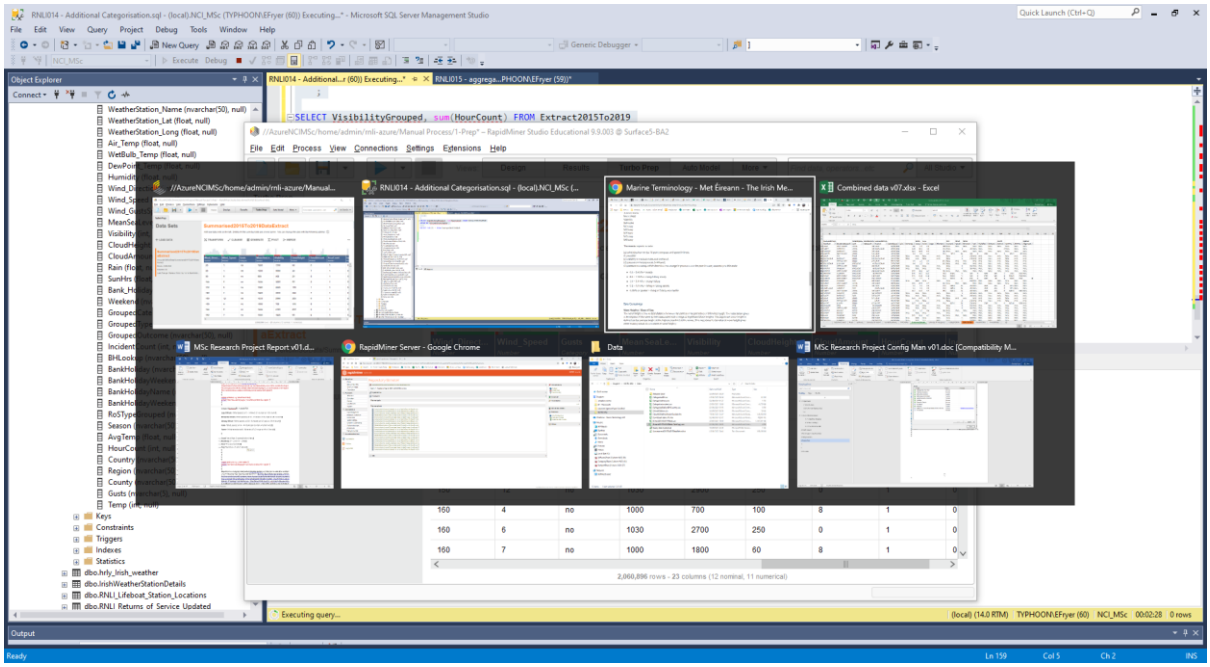
The series of screen prints show how the connection to the sql server was set-up.




SQL Server Import and Export Wizard
Provide a Source Query
Type the SQL statement that will select data from the source database.
SQL statement:
SELECT [YearInteger]
,[HourGroupString]
,[Air_Temp]
,[WetBulb_Temp]
,[DewPoint_Temp]
,[Humidity]
,[Wind_Direction]
,[Wind_Speed]
,[Wind_GustsSpeed]
,[MeanSeaLevelPressure]
,[Visibility]
,[CloudHeight]
,[CloudAmount]
,[Rain]
,[SunHrs]
,[Weekend]
,[GroupedCategory]
,[GroupedType]
,[GroupedOutcome]
,[IncidentCount]
,[BankHoliday]
,[BankHolidayWeekend]
,[BankHolidayName]
,[BankHolidayWeekendName]
,[RoSTypeGrouped]
,[Season]
FROM [NCI_MSc].[dbo].[Extract2015To2019]
Parse
Browse...
Help
< Back
Next >
Finish >>|
Cancel

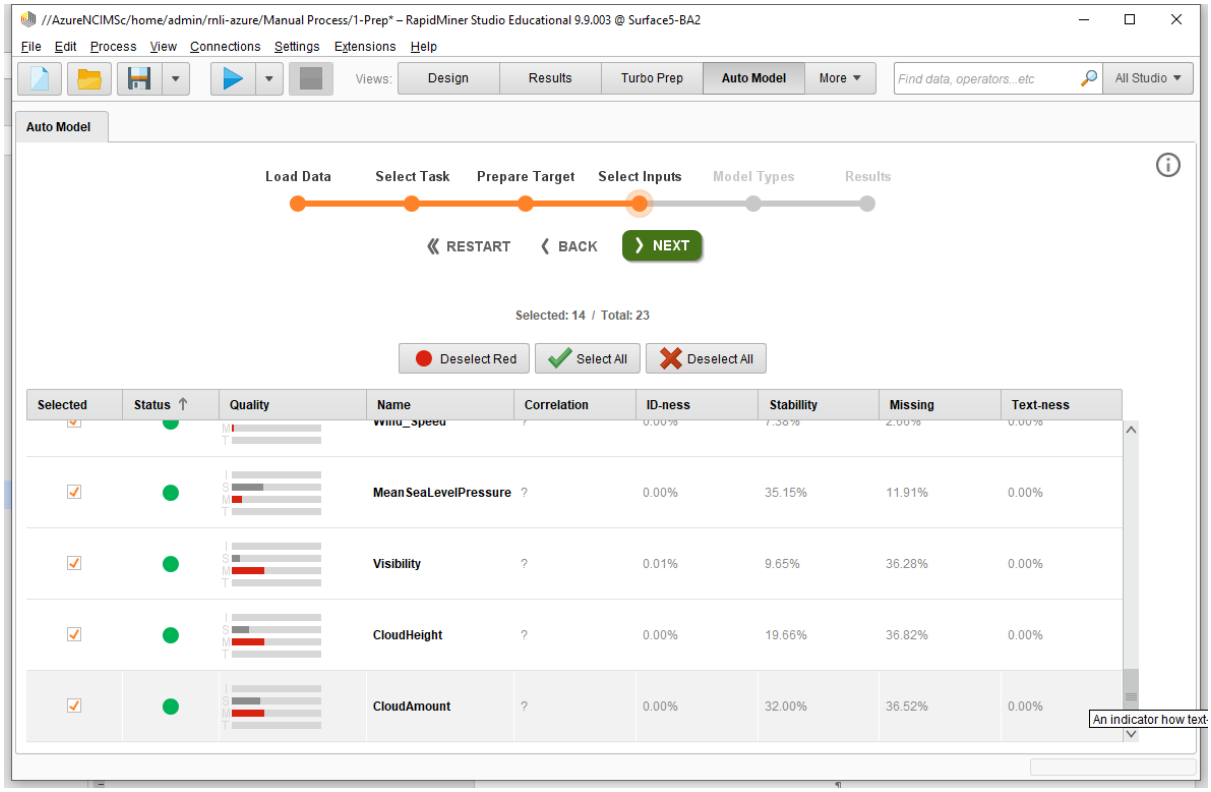

SQL Server Import and Export Wizard
Configure Flat File Destination
Source query:
[Query]
Specify the characters that delimit the destination file:
Row delimiter:
{CR}{LF}
Column delimiter:
Comma {,}
Edit Mappings...
Preview...
Help
< Back
Next >
Finish >>|
Cancel

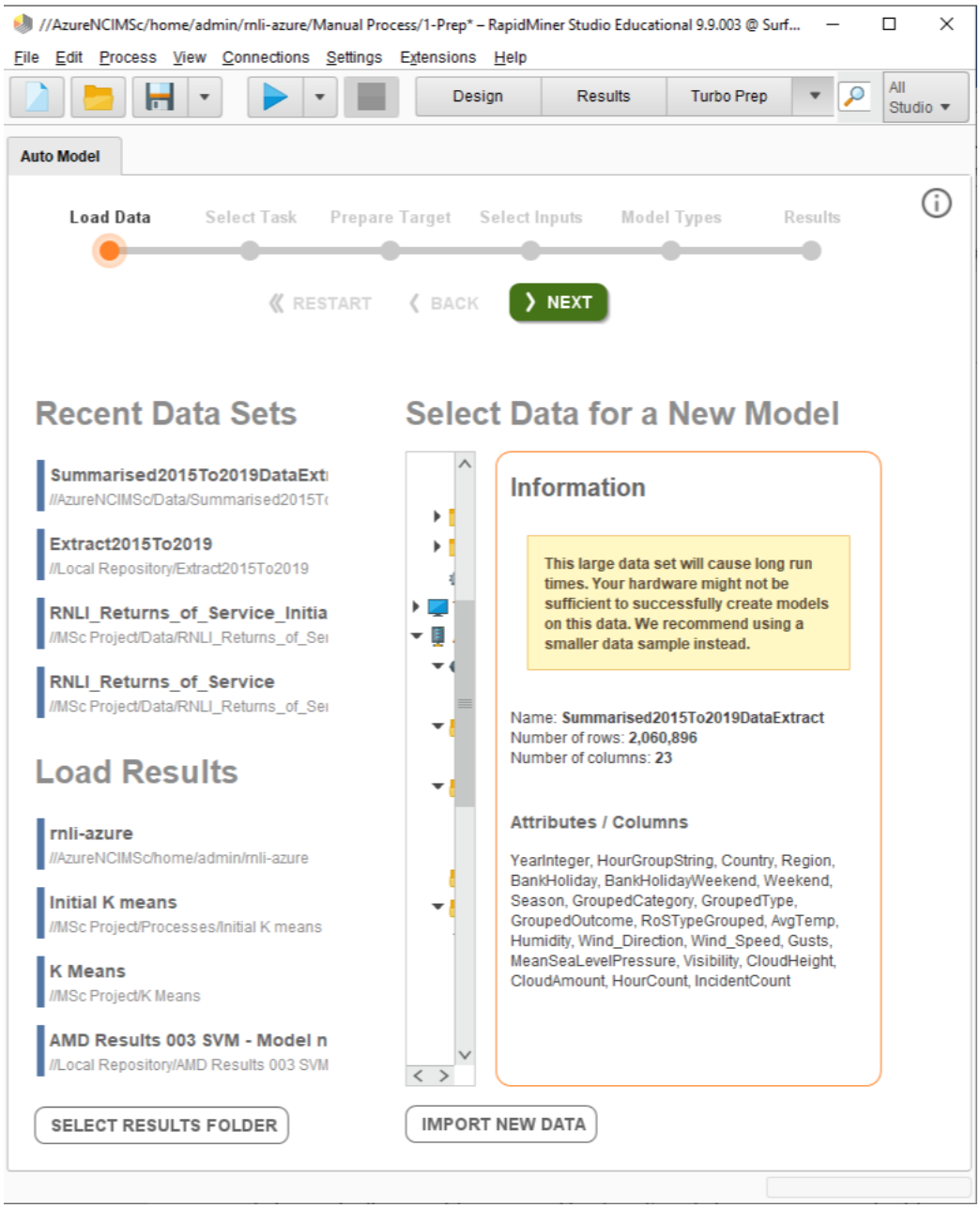

SQL Server Import and Export Wizard
The execution was successful
Success
11 Total
11 Success
0 Error
0 Warning
Details:
Action
Status
Message
Initializing Data Flow Task
Success
Initializing Connections
Success
Setting SQL Command
Success
Setting Source Connection
Success
Setting Destination Connection
Success
Validating
Success
Prepare for Execute
Success
Pre-execute
Success
Executing
Success
Copying to C:\Users\EFryer\Dropbox\NCIRL MSc\Da...
Success
10430112 rows transferred
Post-execute
Success
Stop
Report
Close



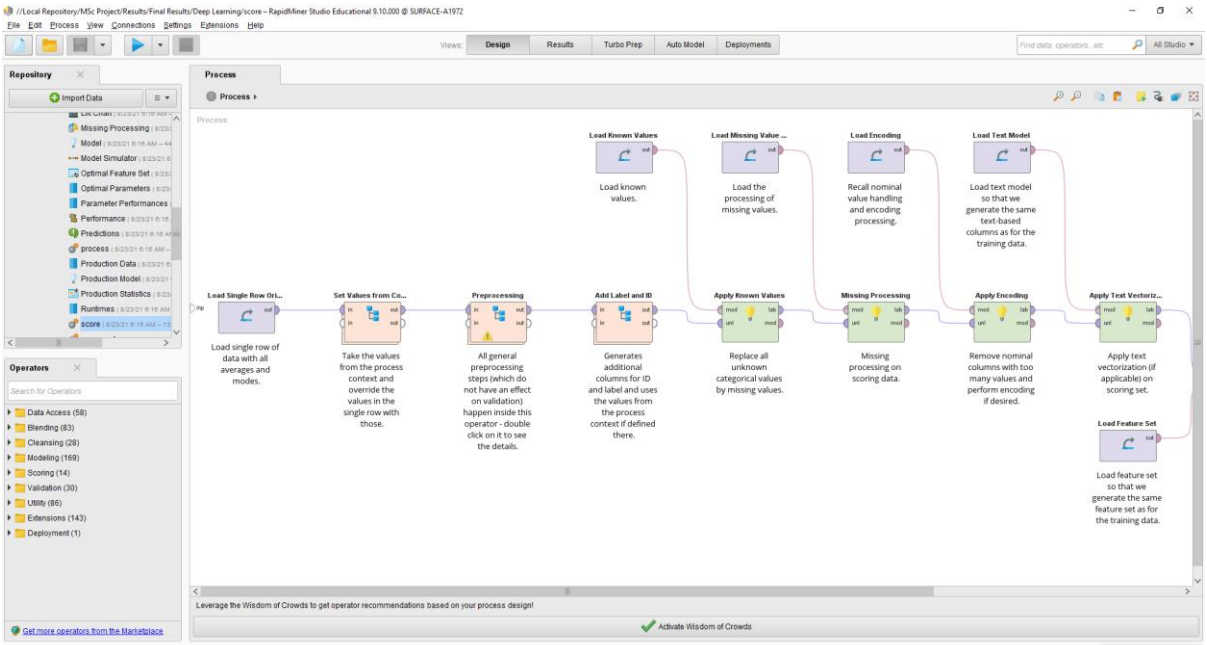
Then connect the server to an Auto model.

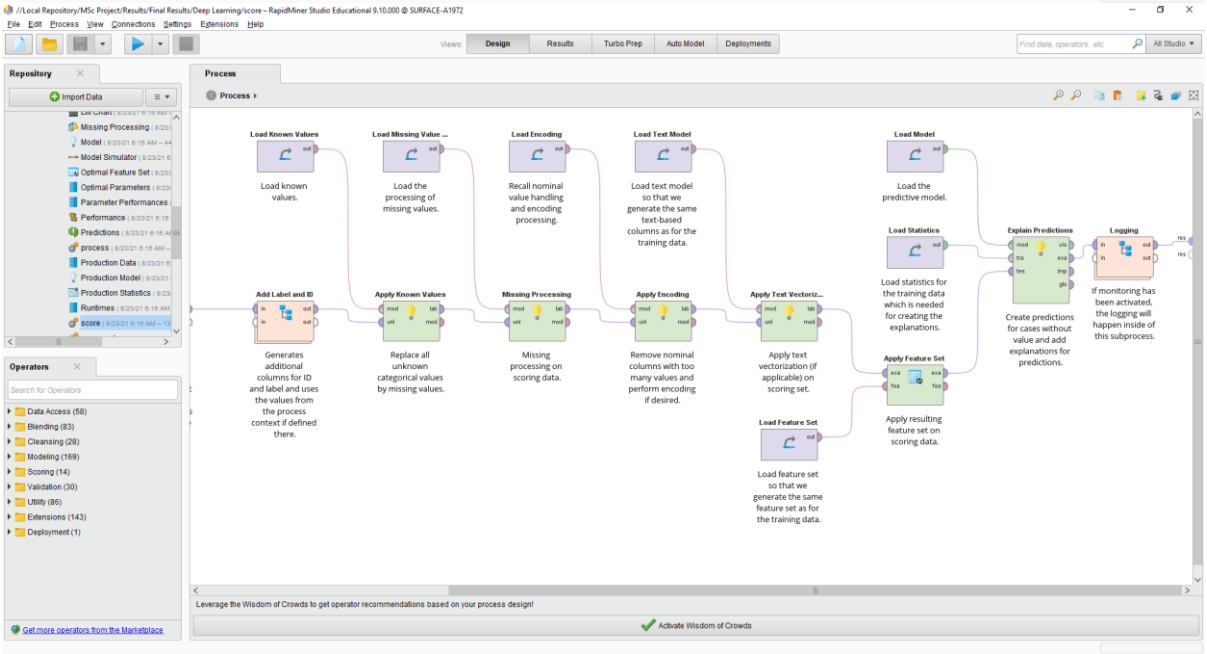
Run various analysis to determine the impact of various variables on the models.




//AzureNCIMSc/home/admin/rnli-azure/Manual Process/1-Prep* – RapidMiner Studio Educational 9.9.003 @ Surf...
File Edit Process View Connections Settings Extensions Help
Design
Results
Turbo Prep
All Studio
Auto Model
Load Data
Select Task
Prepare Target
Select Inputs
Model Types
Results
RESTART
BACK
NEXT
Recent Data Sets
Summarised2015To2019DataExt
//AzureNCIMSc/Data/Summarised2015T
Extract2015To2019
//Local Repository/Extract2015To2019
RNLI_Returns_of_Service_Initia
//MSc Project/Data/RNLI_Returns_of_Ser
RNLI_Returns_of_Service
//MSc Project/Data/RNLI_Returns_of_Ser
Load Results
rnli-azure
//AzureNCIMSc/home/admin/rnli-azure
Initial K means
//MSc Project/Processes/Initial K means
K Means
//MSc Project/K Means
AMD Results 003 SVM - Model n
//Local Repository/AMD Results 003 SVM
SELECT RESULTS FOLDER
Select Data for a New Model
Information
This large data set will cause long run times. Your hardware might not be sufficient to successfully create models on this data. We recommend using a smaller data sample instead.
Name: Summarised2015To2019DataExtract
Number of rows: 2,060,896
Number of columns: 23
Attributes / Columns
YearInteger, HourGroupString, Country, Region, BankHoliday, BankHolidayWeekend, Weekend, Season, GroupedCategory, GroupedType, GroupedOutcome, RoSTypeGrouped, AvgTemp, Humidity, Wind_Direction, Wind_Speed, Gusts, MeanSeaLevelPressure, Visibility, CloudHeight, CloudAmount, HourCount, IncidentCount
IMPORT NEW DATA

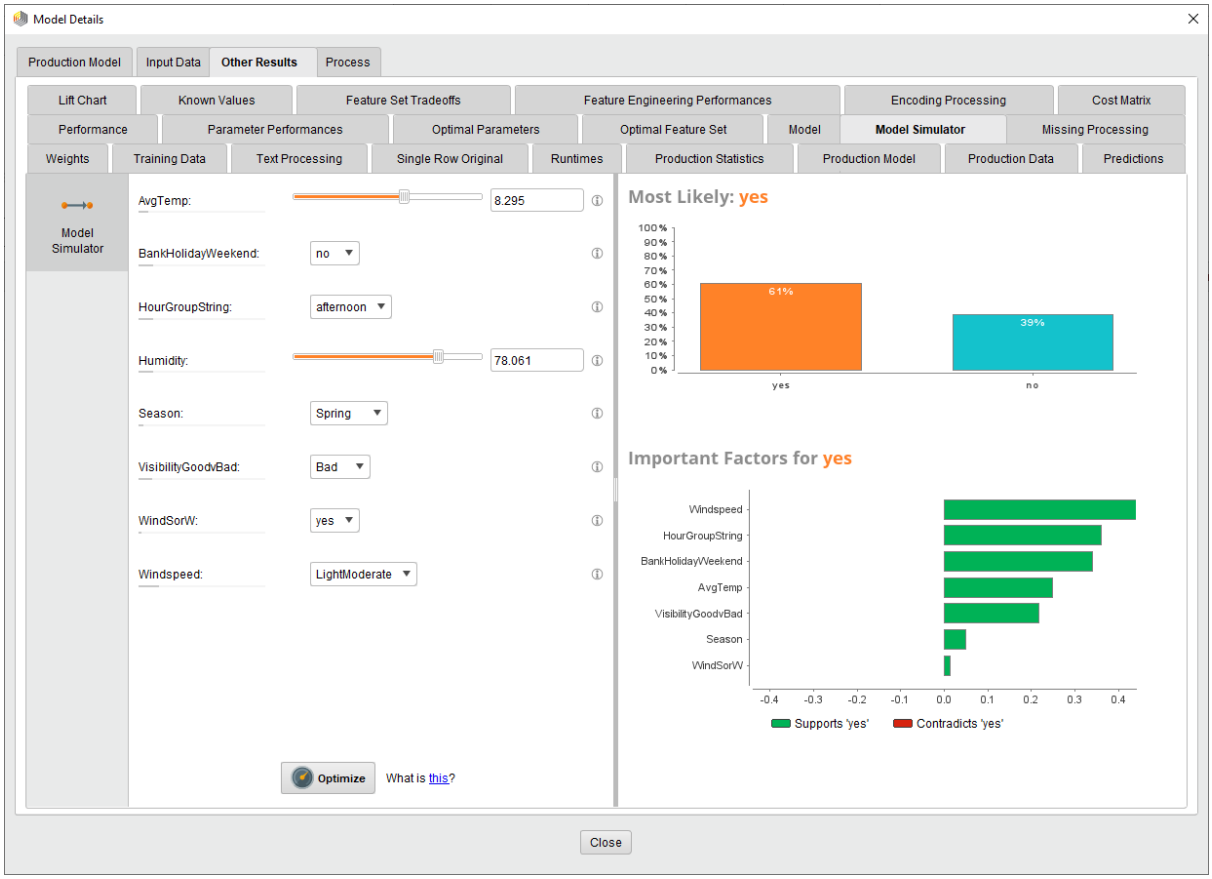
4.1 Examples of models

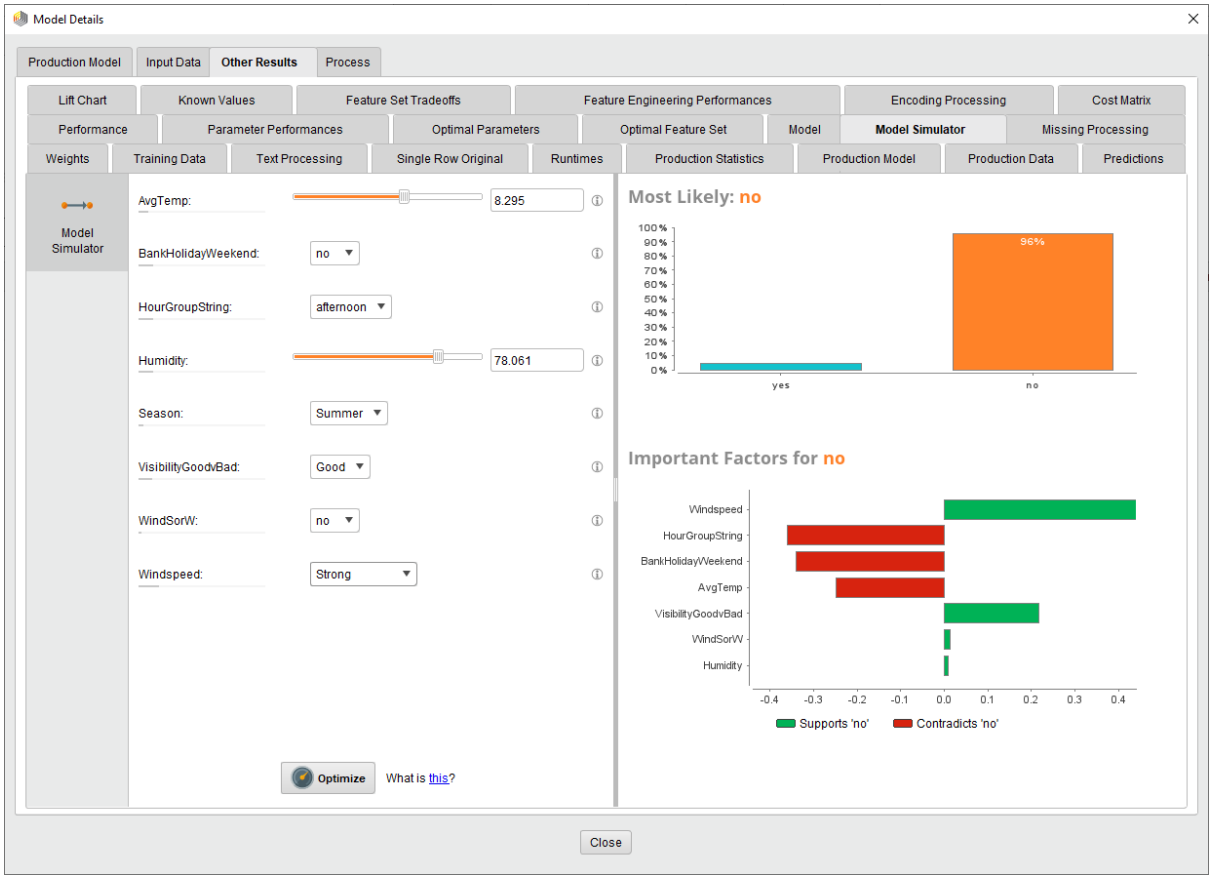




5 Deployment

After evaluating the results for the various models the final models selected can be deployed using the RapidMiner deployment functionality.
This allows the use of the model simulator to predict the outcome for a specific set of variables.




Model Details
Production Model
Input Data
Other Results
Process
Lift Chart
Known Values
Feature Set Tradeoffs
Feature Engineering Performances
Encoding Processing
Cost Matrix
Performance
Parameter Performances
Optimal Parameters
Optimal Feature Set
Model
Model Simulator
Missing Processing
Weights
Training Data
Text Processing
Single Row Original
Runtimes
Production Statistics
Production Model
Production Data
Predictions
Model Simulator
AvgTemp:
8.295
BankHolidayWeekend:
no
HourGroupString:
afternoon
Humidity:
78.061
Season:
Summer
VisibilityGoodvBad:
Good
WindSorW:
no
Windspeed:
Strong
Optimize
What is this?
Most Likely: no
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
96%
yes
no
Important Factors for no
Windspeed
HourGroupString
BankHolidayWeekend
AvgTemp
VisibilityGoodvBad
WindSorW
Humidity
-0.4
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
Supports 'no'
Contradicts 'no'
Close