

SQL Server Machine Learning

Rhode Island SQL User Group
April 3, 2019

John Flannery

Independent Data Scientist / Architect

✉ jflanner@comcast.net

 <http://www.linkedin.com/in/johnwflannery>

 [@AgileDataArch](https://twitter.com/AgileDataArch)



Shameless Self Promotion:

- Independent Data Scientist / Architect available for project work.
- Adjunct Professor of Computer Science – Quinnipiac University
- 39 Years Experience beginning with PL/1 applications accessing IMS databases.
(Translation: I'm old. ☺)
 - CATIC
 - WEX Health (Formally Evolution Benefits, Evolution 1)
 - Travelers (Citi Group)
 - Textron Lycoming
 - The Hartford
- Microsoft Professional Program Certificate in Data Science (October 2018) (Working on AI – July 2019)
- Masters of Science – Computer Science 1995 Rensselaer.
- Past President: Hartford SQL Server User Group.

Our Hypothetical Problem:

Adventure Works operates a call center which – among other things – handles new customer enrollment.

As part of the new customer signup experience – management wants to expose:

- An estimate of the potential revenue the new customer will generate.
- Is the customer likely to purchase a bicycle?

Both estimates should be based on demographics.





The Adventure Works Privacy Policy:

.... We do not share your information with third parties ...

Translation: Azure Machine Learning is
OUT!!



RevoScaleR History: (and RevoScalePy)

Revolution R: Distribution of R written by Revolution Computing. (Location: New Haven, CT. Now: Revolution Analytics.) This version of R dealt with scalability issues – specifically R memory size limitations.

January 23, 2015 – Microsoft purchases the product.

Product rebranded – and included in Microsoft R Server and SQL Server 2016.

RevoscalePy written by Microsoft – using Revolution R as a model.

Both RevoScaleR and RevoScalePy included in Microsoft Machine Learning Server and SQL Server 2017.

RevoScaleR - Details:

- RevoScaleR is a Microsoft specific distribution of R.
 - You get it from Microsoft – not R-Project or Cran.
 - You CAN NOT `install.packages` (RevoScaleR) into an existing R installation.
 - You CAN `install.packages()` all the packages you are used to using into RevoScaleR.
- RevoScaleR is a super set of R. (It's not like C++ vs C#.)
 - Anything written R Foundation “should” work without modification when ported to RevoScaleR. (So says Bill ... err ... Satya.)
 - Anything written in RevoScaleR “should” be portable without modification to R Foundation as long as you have not used rx functions.



RevoScaleR - Benefits:

Chief Benefit – Scalability!! Specifically:

- It can handle datasets larger than installed memory. (It does this by iterating through chunks of data. We will actually see this.)
- Several other limitations of R are addressed. (Example – randomForest has a categorical limit of 32 levels. rxRandomForest does not.)



RevoScaleR - Liabilities:

- For most machine learning algorithms – there is a RevoScale R alternative. (Example: randomForest vs rxRandomForest) One more decision to make.
- The RevoScaleR functions have their own performance metrics. As with all algorithms – tuning is a unique experience.
- The ability to share your result is compromised. If you use rx functions – you can only share with Microsoft Machine Learning (RevoScaleR) shops.
- Release cycles different. Good stuff from R Foundation may take time to work into RevoScale R. Current Versions: Foundation 3.5.2. RevoScale 3.3.3.

Preparing SQL Server:

1. At installation time (or add features to an existing instance) install “SQL Server Machine Learning Services (in Database)”
2. `sp_configure 'external scripts enabled', 1`
Note: reconfigure does not cut it. The instance must be restarted.
(Restarting the instance also starts the new Launcher Service.)
3. Install the MachineLearning schema, and model table. This is custom. What you see is my way of thinking. Only the model column is necessary.



Demo 1:

Preparing SQL and the database

What I did:

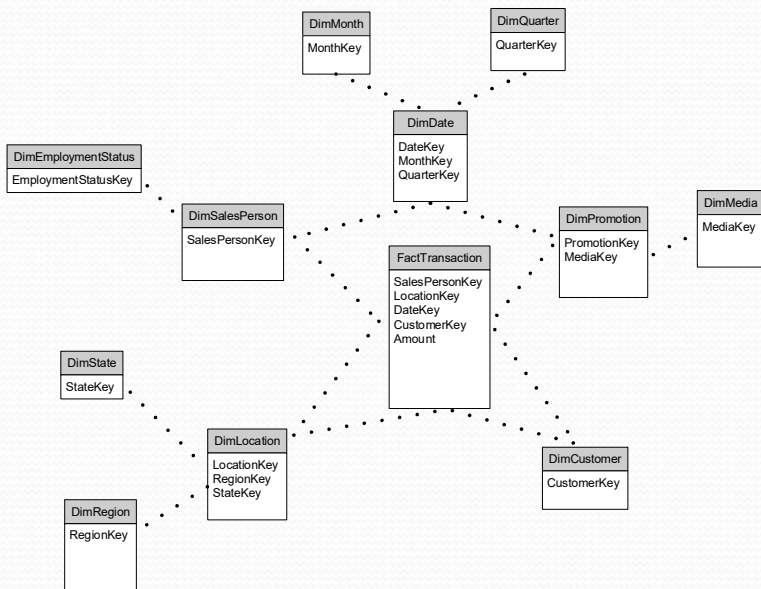
This is a really short demo. I:

- Showed you the sp_configure option.
- Defined the Machine Learning schema in AdventureWorksDW2017.
- Defined the MachineLearning.Models table. I left the demo with the table empty.
- We went into Visual Studio – which was open to the project for this demo. I did a <CTRL>-9 to expose the workspaces menu – showing that I was connected to Microsoft R Server. That will not typically be the case. When you open a new project in VS – you will be connected to your default R installation. You would need to change this early on in your SQL Server Machine Learning project.

Note: I do not talk to Demo slides in the presentation. I include detail of what I intend to do (if the demo Gods are smiling on me) as a prod for myself. But more important – my hope is you will take this material and play with it – and the “What I did” section will provide some guidance.

Issue 1 – Relational Data vs ML Dataset:

SQL Data is typically normalized or
(in our case) STAR schema



Machine Learning wants Fat and Flat.

Size of House	Lot Size (acre)	# of Bedrooms	# of Bathrooms	Price of House
950	2.5	2	1	\$127,325
1,535	1.5	2	2	\$156,570
1,605	2.25	3	1.5	\$158,895
1,905	2.5	2	1.5	\$200,025
2,057	2.25	3	2	\$230,384
2,227	2.75	3	2	\$233,835
3,150	1	4	2	\$261,420
3,620	3	4	3	\$433,500



Solution: Views

- Present data in fat and flat (a.k.a. – Observation) format.
- Make categorical data categorical
 - Translate traditional Flag columns from 1 and 0 to “Yes” and “No”.
- Add columns if doing so can aid Machine Learning.
- Begin Data Exploration. (Look for duplicates and missing values.)



Demo 2:

The views

What I did:

In this demo. I:

- Created and walked through the main view; highlighting the categorical features.
- Began data exploration using that view.
- Created and walked through the Feature and Label views.

Issue 2 – Dataset Schema:

Group Participation Statement:

When you split a dataset into training and testing partitions – data and schema are copied to the new datasets.

Agree



If you choose not to decide – you still have made a choice!!

- Neil Peart

CustomerKey	Age	MaritalStatus	Gender	YearlyIncome	Education	Occupation	HouseOwnerFlag	CommuteDistance	SalesTerritory	PurchasePrice
11000	48	M	M	60000.00	Bachelors	Professional	Yes	1-2 Miles	ST9	Yes
11001	43	S	M	60000.00	Bachelors	Professional	No	0-1 Miles	ST9	Yes
11002	48	M	M	60000.00	Bachelors	Professional	Yes	2-5 Miles	ST9	Yes
11003	46	S	F	70000.00	Bachelors	Professional	No	5-10 Miles	ST9	Yes
11004	40	S	F	60000.00	Bachelors	Professional	Yes	1-2 Miles	ST9	Yes
11005	43	S	M	70000.00	Bachelors	Professional	Yes	5-10 Miles	ST9	Yes
11006	43	S	F	70000.00	Bachelors	Professional	Yes	5-10 Miles	ST9	Yes
11007	50	M	M	60000.00	Bachelors	Professional	Yes	0-1 Miles	ST9	Yes
11008	44	S	F	60000.00	Bachelors	Professional	Yes	10+ Miles	ST9	Yes
11009	50	S	M	70000.00	Bachelors	Professional	No	5-10 Miles	ST9	Yes
11010	50	S	F	70000.00	Bachelors	Professional	No	5-10 Miles	ST9	Yes

CustomerKey	Age	MaritalStatus	Gender	YearlyIncome	Education	Occupation	HouseOwnerFlag	CommuteDistance	SalesTerritory	PurchasePrice
11000	48	M	M	60000.00	Bachelors	Professional	Yes	1-2 Miles	ST9	Yes
11001	43	S	M	60000.00	Bachelors	Professional	No	0-1 Miles	ST9	Yes
11002	48	M	M	60000.00	Bachelors	Professional	Yes	2-5 Miles	ST9	Yes
11003	46	S	F	70000.00	Bachelors	Professional	No	5-10 Miles	ST9	Yes
11004	40	S	F	60000.00	Bachelors	Professional	Yes	1-2 Miles	ST9	Yes
11005	43	S	M	70000.00	Bachelors	Professional	Yes	5-10 Miles	ST9	Yes
11006	43	S	F	70000.00	Bachelors	Professional	Yes	5-10 Miles	ST9	Yes
11007	50	M	M	60000.00	Bachelors	Professional	Yes	0-1 Miles	ST9	Yes

CustomerKey	Age	MaritalStatus	Gender	YearlyIncome	Education	Occupation	HouseOwnerFlag	CommuteDistance	SalesTerritory	PurchasePrice
11008	44	S	F	60000.00	Bachelors	Professional	Yes	10+ Miles	ST9	Yes
11009	50	S	M	70000.00	Bachelors	Professional	No	5-10 Miles	ST9	Yes
11010	50	S	F	70000.00	Bachelors	Professional	No	5-10 Miles	ST9	Yes

BirthDate	date
MaritalStatus	Factor w/ 2 levels: M, S
Gender	Factor w/ 2 levels: F, M
YearlyIncome	int
TotalChildren	int
NumberChildrenAtHome	int
Education	Factor w/ 5 levels: Partial High School, High School, Partial College, Bachelors, Graduate Degree
Occupation	Factor w/ 5 levels: Clerical, Management, Manual, Professional, Skilled Manual
HouseOwnerFlag	Factor w/ 2 levels: No, Yes
NumberCarsOwned	int
CommuteDistance	Factor w/ 5 levels: 0-1 Miles, 1-2 Miles, 2-5 Miles, 5-10 Miles, 10+ Miles
SalesTerritory	Factor w/ 10 levels: ST1, ST10, ST2, ST3, ST4, ST5, ST6, ST7, ST8, ST9

BirthDate	date
MaritalStatus	Factor w/ 2 levels: M, S
Gender	Factor w/ 2 levels: F, M
YearlyIncome	int
TotalChildren	int
NumberChildrenAtHome	int
Education	Factor w/ 5 levels: Partial High School, High School, Partial College, Bachelors, Graduate Degree
Occupation	Factor w/ 5 levels: Clerical, Management, Manual, Professional, Skilled Manual
HouseOwnerFlag	Factor w/ 2 levels: No, Yes
NumberCarsOwned	int
CommuteDistance	Factor w/ 5 levels: 0-1 Miles, 1-2 Miles, 2-5 Miles, 5-10 Miles, 10+ Miles
SalesTerritory	Factor w/ 10 levels: ST1, ST10, ST2, ST3, ST4, ST5, ST6, ST7, ST8, ST9

Issue 2 – Dataset Schema:

Goal – Real Time Scoring

Implication – 1 Observation

Feature Names – exact

Data – Reasonable – conforms with training schema.

Schema – slightly off

CustomerKey	Age	MaritalStatus	Gender	YearlyIncome	Education	Occupation	HouseOwnerFlag	CommuteDistance	SalesTerritory	PurchasedBike
11000	48	M	M	90000.00	Bachelors	Professional	Yes	1-2 Miles	ST9	Yes
11001	43	S	M	60000.00	Bachelors	Professional	No	0-1 Miles	ST9	Yes
11002	48	M	M	60000.00	Bachelors	Professional	Yes	2-5 Miles	ST9	Yes
11003	46	S	F	70000.00	Bachelors	Professional	No	10-15 Miles	CT9	Yes
11004	40	S	F	80000.00	Bachelors	Professional	Yes			
11005	43	S	M	70000.00	Bachelors	Professional	Yes			
11006	43	S	F	70000.00	Bachelors	Professional	Yes			
11007	50	M	M	60000.00	Bachelors	Professional	Yes			
11008	44	S	F	60000.00	Bachelors	Professional	Yes			
11009	50	S	M	70000.00	Bachelors	Professional	No			
11010	50	S	F	70000.00	Bachelors	Professional	No			

BirthDate	date
MaritalStatus	Factor w/ 2 levels: M, S
Gender	Factor w/ 2 levels: F, M
YearlyIncome	int
TotalChildren	int
NumberChildrenAtHome	int
Education	Factor w/ 5 levels: Partial High School, High School, Partial College, Bachelors, Graduate Degree
Occupation	Factor w/ 5 levels: Clerical, Management, Manual, Professional, Skilled Manual
HouseOwnerFlag	Factor w/ 2 levels: No, Yes
NumberCarsOwned	int
CommuteDistance	Factor w/ 5 levels: 0-1 Miles, 1-2 Miles, 2-5 Miles, 5-10 Miles, 10+ Miles
SalesTerritory	Factor w/ 10 levels: ST1, ST10, ST2, ST3, ST4, ST5, ST6, ST7, ST8, ST9

Age	MaritalStatus	Gender	YearlyIncome	Education	Occupation	HouseOwnerFlag	CommuteDistance	SalesTerritory
35	M	M	40000.00	Partial College	Skilled Manual	Yes	1-2 Miles	ST4

BirthDate	date
MaritalStatus	Factor w/ 1 levels: M
Gender	Factor w/ 1 levels: M
YearlyIncome	int
TotalChildren	int
NumberChildrenAtHome	int
Education	Factor w/ 1 levels: Partial College
Occupation	Factor w/ 1 levels: Skilled Manual
HouseOwnerFlag	Factor w/ 1 levels: No
NumberCarsOwned	int
CommuteDistance	Factor w/ 1 levels: 5-10 Miles
SalesTerritory	Factor w/ 1 levels: ST9



No Prediction
for You!!

NEXT!!!



Issue 2 – Schema Management:

Issue:

- We have blocks of code (like schema declaration) that will be repeated often.
- As we go on in this journey – we are probably going to come upon neat things specific to our business.
- One of the big benefits of R – Sharing!! This is not kindergarten.

Resolution: Lets make our own R Library!!

Issue 3 – How are we going to do Library Management?:

Issue:

- When you (the developer) do an `install.package` – the package is installed in your personal library.
- Yours is not the ID running SQL Server.
- That personal Library will not be in the security scope of SQL Server.

Resolution: There is going to need to be an SDLC process to keep SQL Server libraries up to date. Mechanical steps needed in that process are included in the next demo.



Demo 3:

Library Management - Setting up the AdventureWorks library

What I did:

In this demo. I:

- Walked you through the AdventureWorks library definition – highlighting the MAN and R files. (To add more functions to this library – you simply add man and r files.)
- Executed the RGUI that came with Microsoft Machine Learning Server. RUN AS ADMINISTRATOR. Ran the install package into both the Machine Learning Server library and the library servicing my instance.
- Showed you the Johns Packages script – which you need to run if you want to set this up at home. Highlighted the fact there are two sections in this script.



The process – Azure ML Studio like:

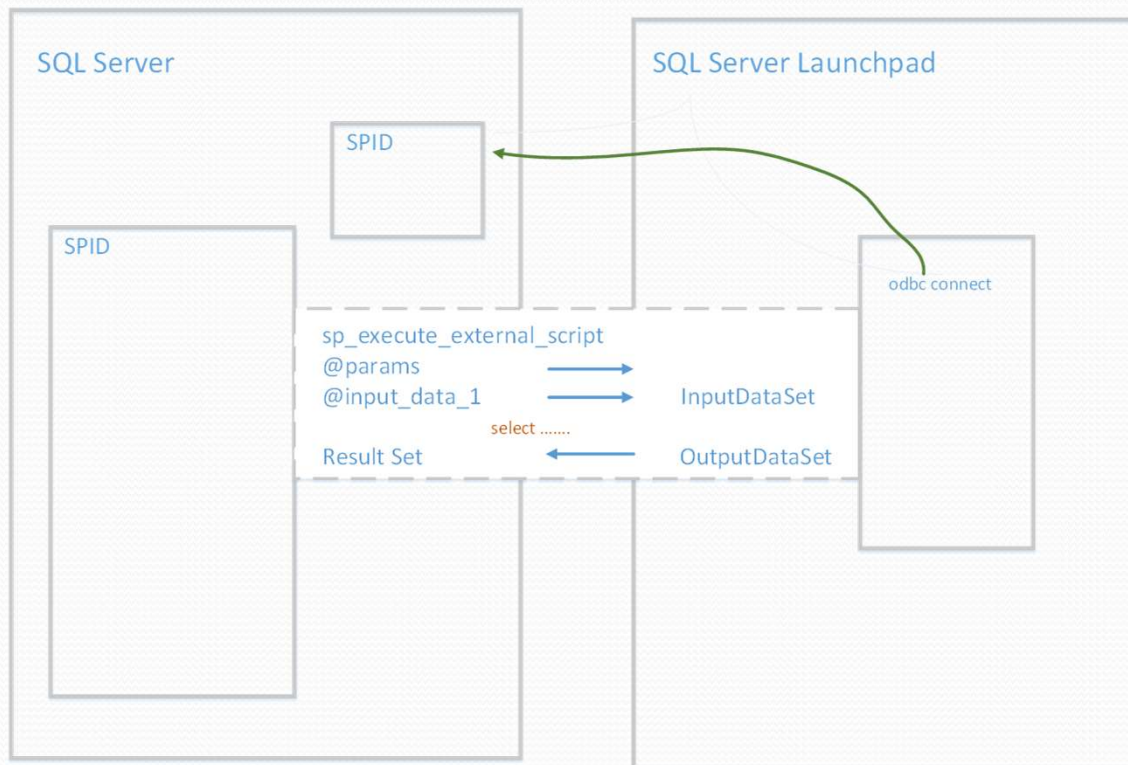
Azure Machine Learning Studio

- Create a Training experiment.
 - Explore Data
- Create a Predictive Experiment
- Publish a Web Service

SQL Server Machine Learning

- Create a Training experiment.
 - Explore Data
(Visual Studio)
- Create a Predictive Model
(Stored Procedure)
- Create a Prediction Stored Procedure.

Internals:



For complete zip file – email jflanner@comcast.net
That is not a typo. I am old. There was a time email addresses were restricted to 8 characters.

SQL Server Launchpad – This service runs the advanced analytics.

sp_execute_external_script is bridge between services.

3 parameters (white lie):

- **@params** Used to define other scalar parameters.
- **@input_data_1** is the **only** dataset transfer between SQL and Launchpad.
 - MUST be a select statement to be executed on the Launchpad side. (No select * from @MyTable.)
 - data_1 does not imply data_2 exists.
- **Output Data Set** is the **only** dataset transfer from Launchpad back to SQL.

Once in Launchpad – your R script can open an ODBC connection back to SQL.

The original spid waits while R is running.



Demo 4:

The Big Demo!!

In this demo. I:

- Walked through the two models in Visual Studio. Highlighted:
 - The TrainedModel variable.
 - The RowsRead chunk thing. How RevoScale R processes datasets.
 - The distribution of revenue. (Which is not normal.)
- The Publication stored procedure. Highlighted:
 - Subset of the R code copied directly into the sproc.
 - How the trained model is returned to the sproc and stored in the model table.
 - The dataset -> Result Set interface.
- The Prediction Stored Procedure. Highlighted:
 - The dynamic SQL used to build the one observation dataset.
 - How the Model and Prediction interface between the spid and Launch Pad

What is up with that “1.0 – score” thing??

“Yes” and “No” are just words – or Labels – in the Bike Buyers model. I could have labeled the alternatives “Fred” and “Ginger”. Alphabetically (the order default for categorical) “No” comes before “Yes.”


RxGlm does not compute probability. It simply plots a point (the score) between two alternatives using a 0 – 1 interval.

confusionMatrix considers the alternative on the left – what ever you label it – to be the “Positive Alternative.” For this reason – I relabeled BikeBuyer “Yes” and “No” – the opposite of the alphabetical default.

The alternative “on the left” is also on the 0 side of the interval.



Let's have a Sale!!!



Click image to open expanded view

List Price: ~~\$25.00~~
Price: **\$11.40** ✓prime
You Save: **\$13.60 (54%)**

Get \$70 off instantly: Pay \$0.00 upon approval for the Amazon Prime Rewards Visa Card.

Note: Available at a lower price from [other sellers](#), potentially without free Prime shipping.

FREE Delivery by **Saturday**
if you order within 8 hrs 7 mins, or

Get it **Tomorrow** if you order within 1 hr 37 mins and choose paid shipping at checkout. [Details](#)
In Stock.





Ships from and sold by Amazon.com. Gift-wrap available.

Size: **1-Pack**

1-Pack \$11.40 ✓prime	2-Pack \$23.02	4-Pack \$38.31	12-Pack \$117.14
---	-------------------	-------------------	---------------------

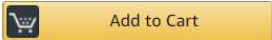
- Clears cloudy spa water fast
- Compatible with chlorine, bromine, ozone and biguanide sanitizers
- For more than 25 years, Leisure Time has been the premier name in spa water care
- Helps spa filters perform at peak efficiency
- Compatible with chlorine, bromine, ozone and biguanide sanitizers
- Utilizes polymer action to remove suspended particles


New (37) from \$11.35 ✓prime


Share    

☒ **One-time purchase:**
\$11.40

Qty: 1 ▾

 Add to Cart

 Buy Now

 Deliver to John - Burlington 06013

☐ **Subscribe & Save:**
\$11.40

Add to List ▾

[Add to your Dash Buttons](#)

Can you spot the problem here????



The Suckah Sale:

- We are going to “sell” our most expensive bike at 2x normal cost.
- Offer this sale to all our current customers.
- Experience 75% conversion. It’s a beautiful thing.
- Can I make stuff up or what!!!



Demo 5:

The Suckah Sale!!

What I did:

In this demo. I:

- Ran the update script demonstrating the sale.
- Showed that the views have adjusted. But the models have not.
- Installed the job to run the training stored procedures. Ran the job after installation.
- Demonstrated that the models have now adjusted.

The Rat Trap I want down – Zip Code:

Adventure Works maintains PostalCode as part of address.

ZipCode defines a chunk of land. 06013 describes Burlington CT.

It is reasonable to assume that:

- Some portion of the population of Burlington owns bicycles.
- Some portion of the population of 06443 (Madison, CT) own bicycles.
- The two are not equal. Burlington has hills, Madison has beaches. Burlington has Bike Trails. Madison has traffic.

From a Machine Learning perspective – I thought this feature would be interesting.





Not

Way to many of them.

Lots of literature suggests relating zip code to:

- Population Density
- Average Income
- Rain Days per Year
- Average High Temperature

This information is available from data.gov.

<https://catalog.data.gov/dataset/demographic-statistics-by-zip-code-acfc9>

How to set up this scenario at Home...

- Make sure you have SQL Server Machine Learning installed on your machine.
- You obviously have this project downloaded. Put the R project in an appropriate project area and the SQL where appropriate. (An SSDT project for example. 😊)
- Download and recover AdventureWorks 2017DW from <https://docs.microsoft.com/en-us/sql/samples/adventureworks-install-configure?view=sql-server-2017>
- Install Visual Studio R tools. <https://docs.microsoft.com/en-us/visualstudio/rtvs/installing-r-tools-for-visual-studio?view=vs-2017>
- Make sure the R workspace in Visual Studio is pointed at Microsoft R server.
- At this point – just run through the scripts in the demos.



References...

- Really good article on R and SQL Server
<https://stephanefrechette.com/data-analytics-r-sql-server/#.XDDbLVxKiUk>
- Creating an R library from scratch.
<https://hilaryparker.com/2014/04/29/writing-an-r-package-from-scratch/>
- Implementing Advanced Analytics with SQL Server 2017 and Python
<https://www.pass.org/24hours/2017/summitpreview/Sessions/Details.aspx?sid=66937>



Thank You!!!!!!