A Survey of Current End-user Data Analytics Tool Support

Hourieh Khalajzadeh, Deakin

Mohamed abdelrazek, Deakin

John Grundy, Monash

John Hosking, Auckland

Qiang He, Swinburne

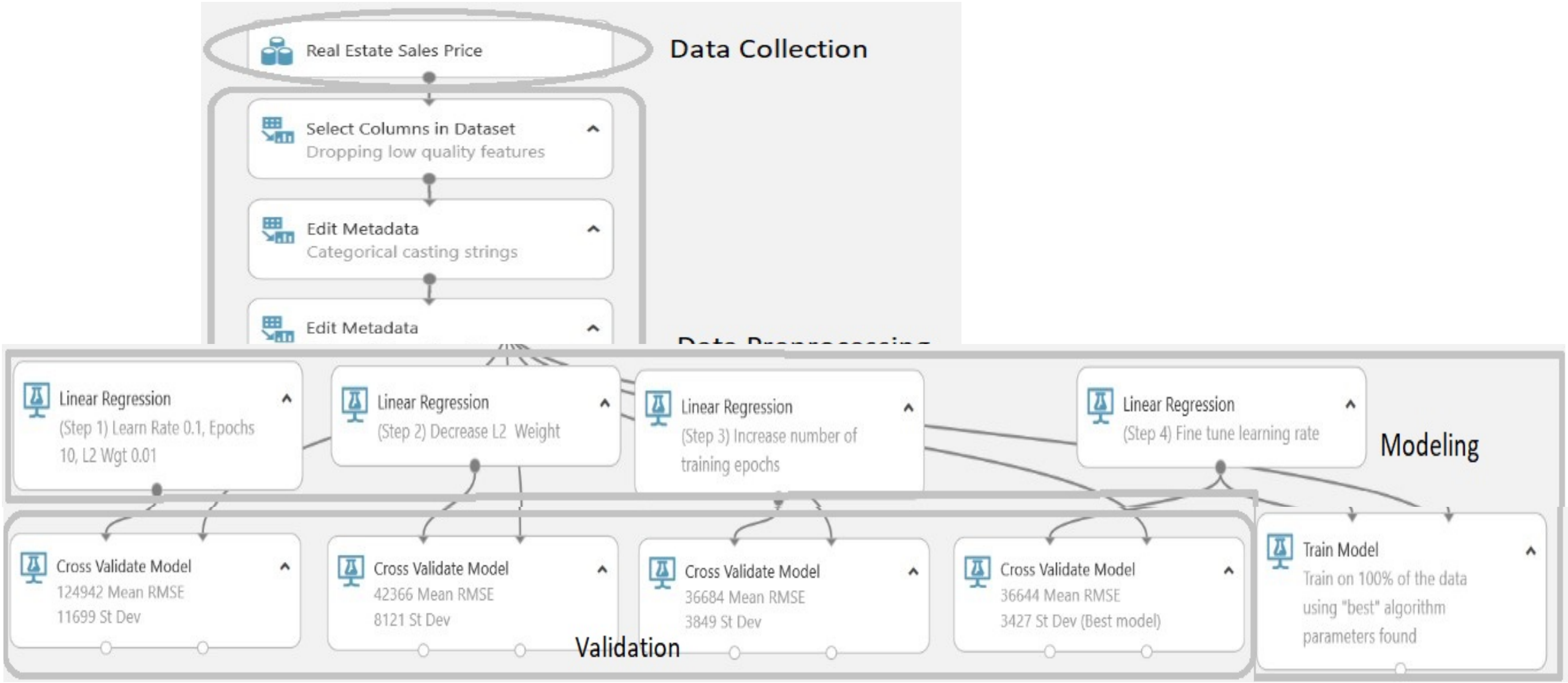
Outline

- Data analytics stages
- Key requirements for end user data analytics tools
- Existing end user data analytics tools
- Issues, strengths and weaknesses
- Research directions
- Conclusions

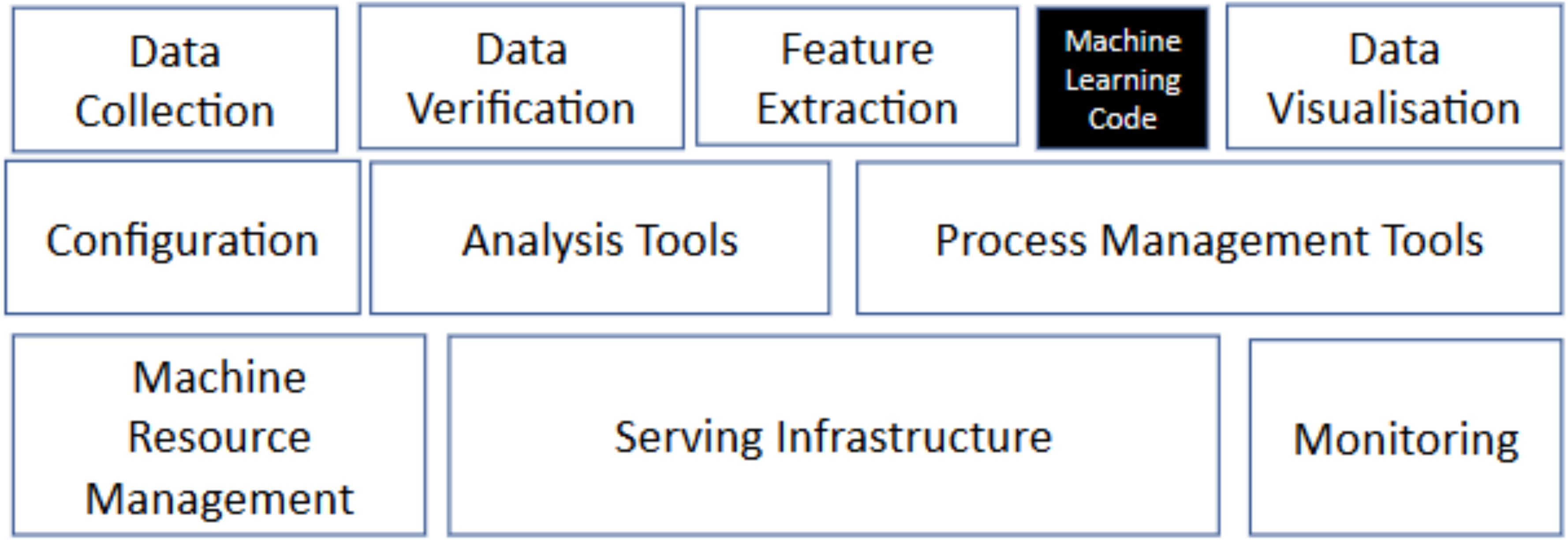
- Classifying the problem
- Acquiring data
- Processing data
- Modeling the problem
- Validation and execution
- Deploying

C. E. Sapp, "Preparing and Architecting for Machine Learning", Gartner Technical Professional Advice, 2017.

Example: Real Estate Sales Price Prediction Project in Azure ML Studio



Artificial Intelligence Systems Development Building Blocks



Only a small component of real-world ML systems is the ML model.
The required surrounding infrastructure is vast and complex.

- Elicitation & Analysis of the requirements
- Design
- Implementation
- Testing
- Maintenance

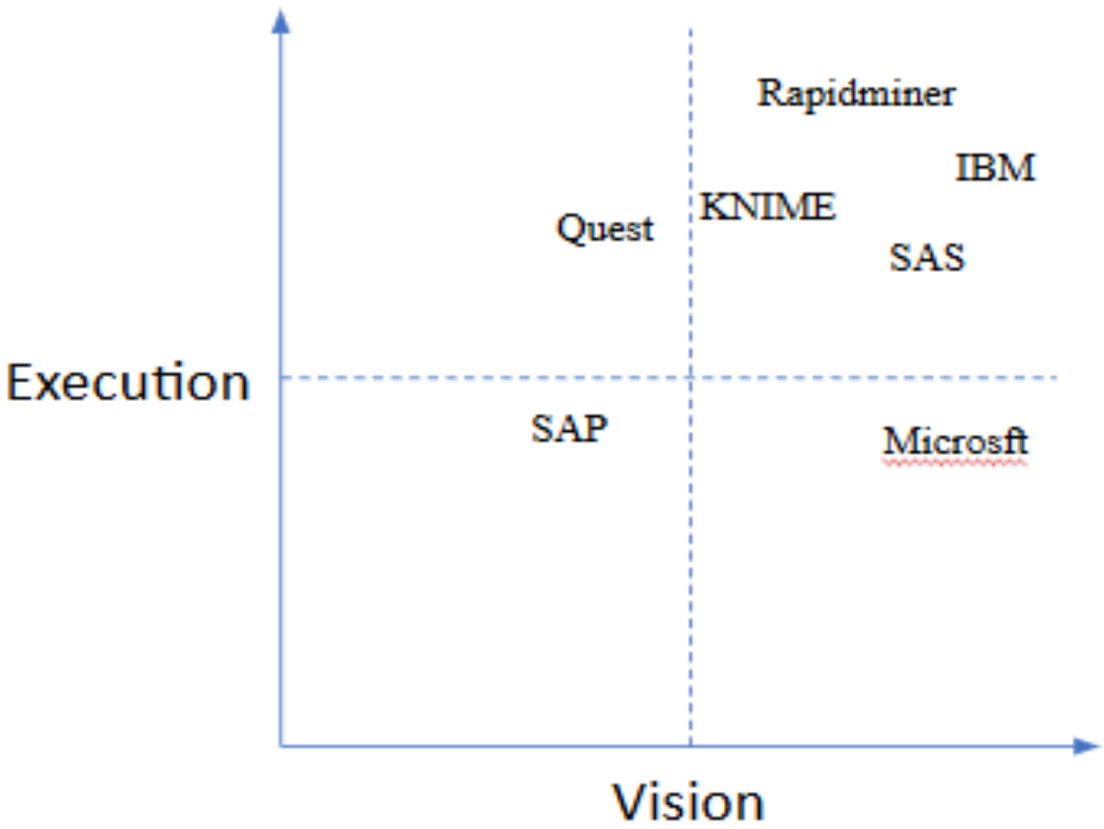
What should Big Data Analytics Software support?

- Diverse data ingestion
- Wrangling and cleansing
- Data integration and querying for very large data volumes
- Feature extraction and selection
- Tailoring and combination of diverse data analytics techniques
- Integration of diverse software and services
- Communication of findings and integration with existing IT solutions
- Quality of service attributes including: scalability, privacy, security, reliability and adaptability to changes in the target environment

Key Requirements for End User Data Analytics Tools

- Support all data preprocessing operations e.g. cleaning, wrangling, anomaly detection
- Want it to be understandable and useable for domain experts, data scientists, and even users with very limited data science and programming knowledge
- Cover a variety of the algorithms for each stage of data processing, modeling and evaluation processes.
- Offer flexible options for experienced users such as data scientists
- Cover all AI-SDLC stages including problem description, requirements, design, implementation, testing and deployment
- Be industry ready for large scale industry-based projects
- Be cost effective, be deployable on the cloud, on premises or both

Gartner 2017 Magic Quadrant for Data Science Platforms

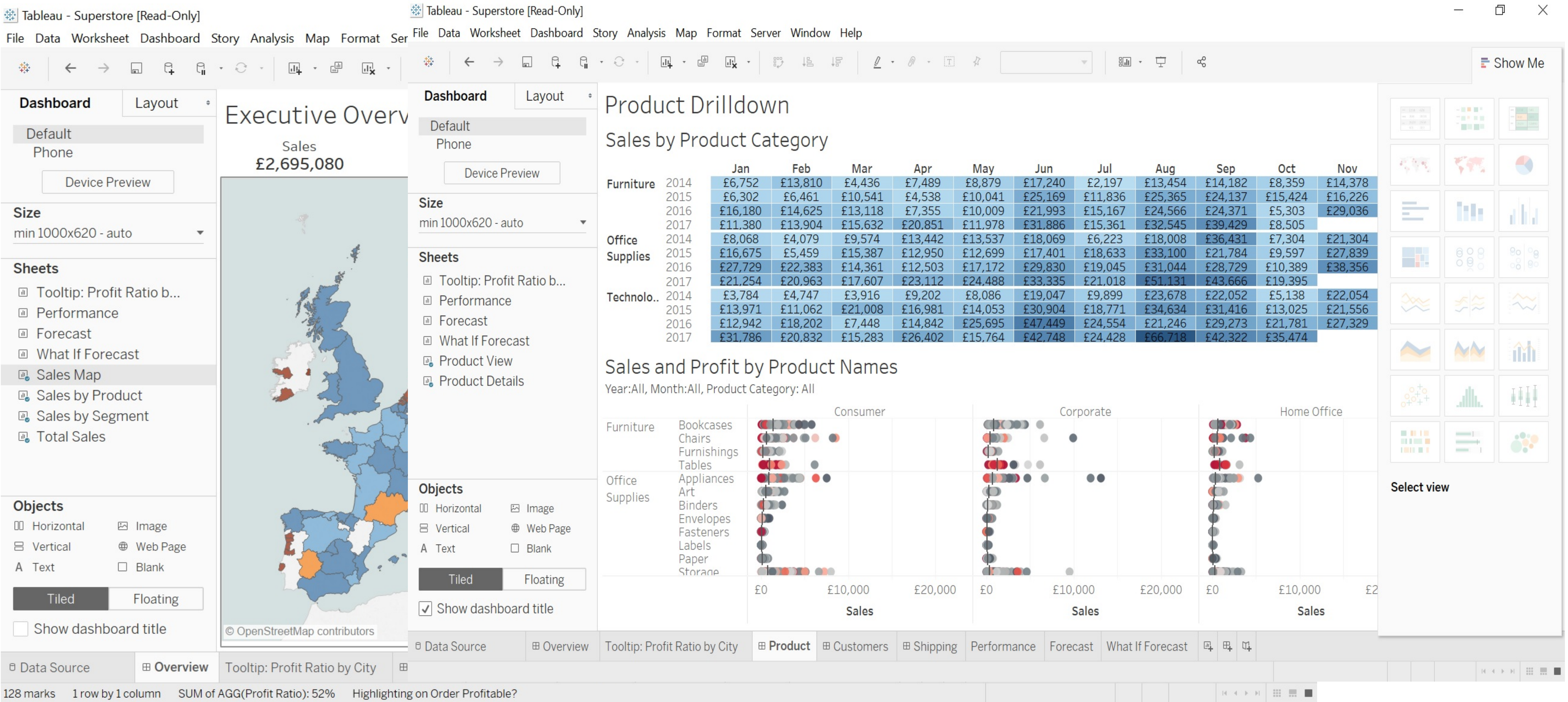


Existing Data Analytics Tools & AI Systems Building Blocks

- Variety of tools developed to automate the ML code as well as the data verification and feature extraction phases
- We group these components (building blocks of an AI-powered systems) into three groups:
  - **DataOps -** includes data collection/ingestion, data validation cleansing, wrangling, filtering, union, merge, etc.
  - **AIOps** - covers feature engineering and model selection, model training and tuning, use of variety of ML, AI techniques
  - **DevOps** - covers model integration and deployment, monitoring and serving infrastructure

- tools such as Tableau, Plotly, and Trifacta
- focus on data operations such as visualization, data cleaning, data wrangling, and so on.
- Interactive visualisation

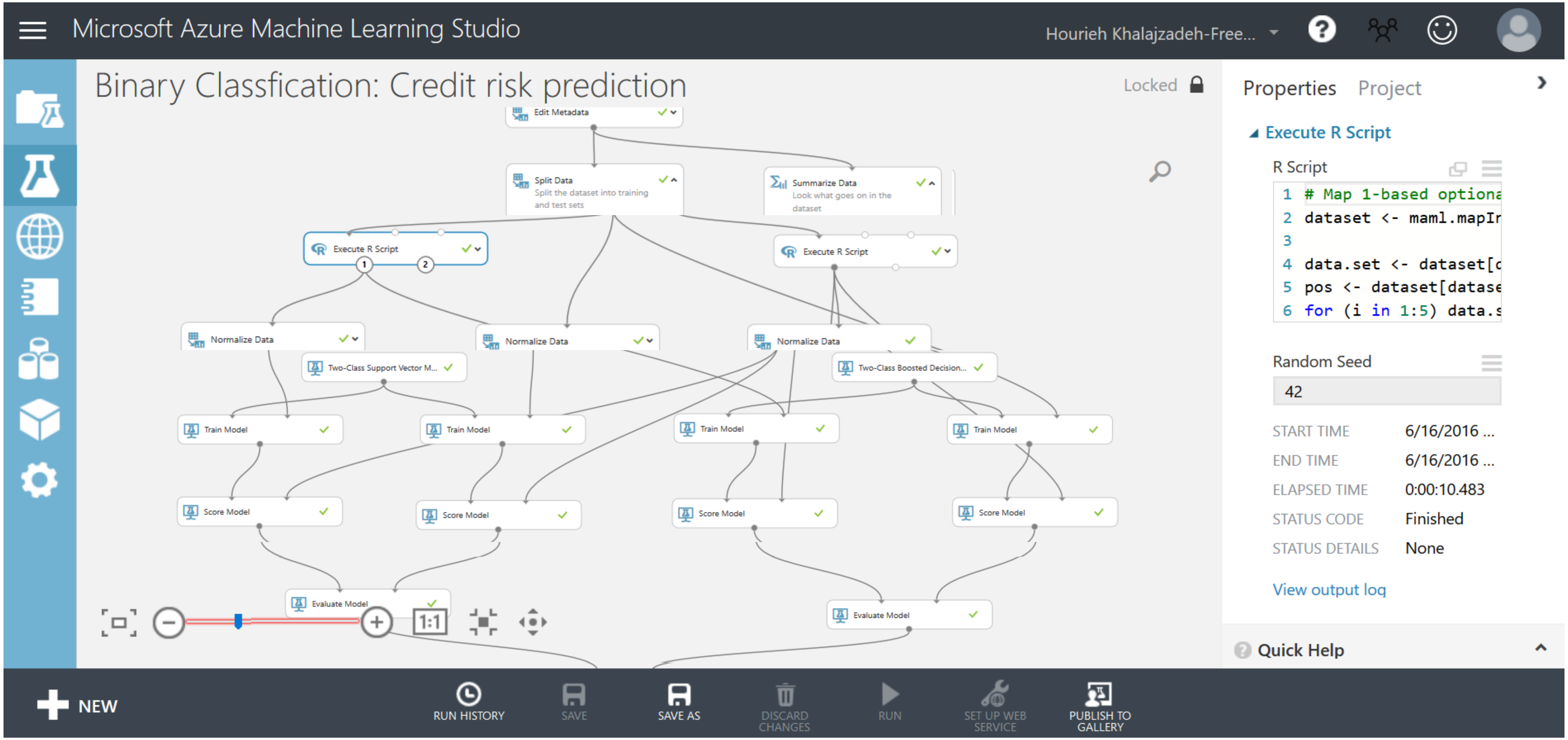
Example of Tableau in use for real estate data analysis



From the Viewpoint of AIOps

- large number of tools focusing on the artificial intelligence and machine learning operations
- Some examples are Azure ML Studio, Amazon AWS ML, Google Cloud ML, BigMl, Weka, Rapidminer, IBM Watson ML, SAS, KNIME, and Tensorport
- tools in this group also often cover DataOps to some extent

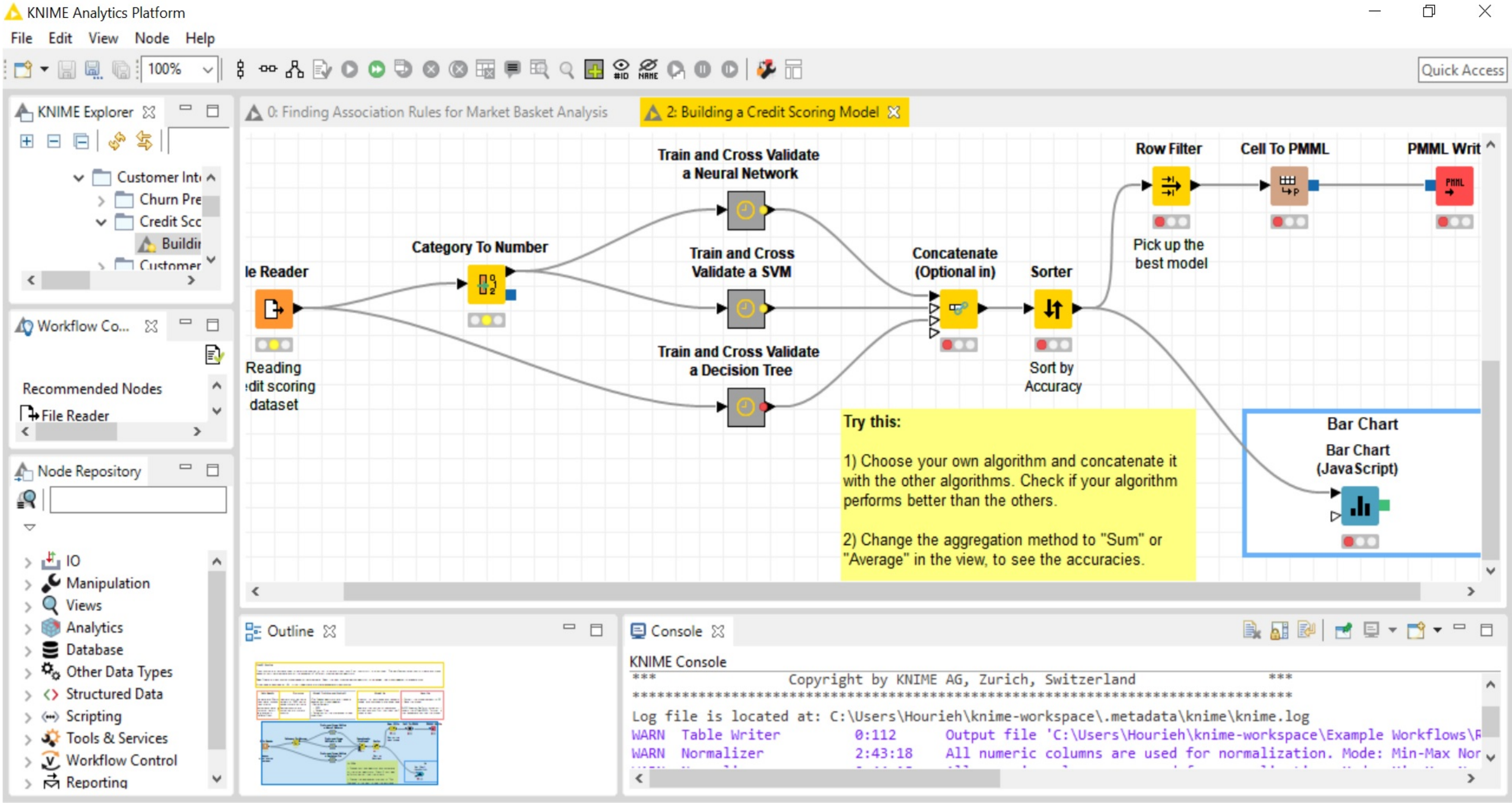
An example of Azure ML Studio in use



From the DevOps Point of View

- Some tools focus on the deployment of the solutions on the cloud or on premises as well as building industry ready solutions
- Some examples are Rapidminer, IBM Watson ML, SAS, and KNIME
- These tools assist to prepare industry ready solutions deployable on both cloud and on premises

An example of KNIME in use



Strengths and Weaknesses (see paper for details…)

End Users Tools	SDLC phases								Tool usability						
	Business problem description	Requirements	Design	Implementation		Testing	Deployment								
							DevOps			Cost					
				DataOps	AIOps		Cloud based	On premises	Industry ready	Free trial/for limited access	Plan based/pay as you go	Usability	Comprehensiveness	Flexibility	No Data science knowledge required
Tableau				✓			✓	✓	✓	✓	✓	✓	✓		✓
Plotly				✓			✓	✓		✓	✓		✓	✓	
Trifacta				✓			✓			✓		✓			✓
Azure ML Studio				✓	✓		✓			✓	✓	✓		✓	
Amazon AWS ML				✓	✓		✓				✓	✓			
Google Cloud ML				✓	✓		✓			✓	✓		✓		
BigML				✓	✓		✓	✓		✓	✓	✓		✓	
Weka				✓	✓			✓		✓		✓			
Rapidminer				✓	✓		✓		✓	✓	✓				
IBM Watson ML				✓	✓		✓	✓	✓		✓	✓	✓	✓	
SAS				✓	✓		✓	✓	✓		✓		✓	✓	
KNIME				✓	✓		✓		✓	✓		✓	✓	✓	
TensorPort					✓			✓	✓	✓			✓	✓	

Gaps in existing tools. (see paper for details)

- Current practices and tools do not cover most activities of analysis and design, esp business requirements
- Most focus on low-level data analytics process design, coding and visualization of results
- Most assume data is in a form amendable to processing – but most datasets are not "clean" nor "integrated", and great effort is needed to source the data, integrate, cleanse, harmonize, pre-process it
- Only a few offer the ability for data science experts to embed new code and expand the algorithms based on their needs
- Most only cover parts of the DataOps, AIOps, and DevOps of the data analytics life cycle
- Many real-world problems require large datasets to be processed and thus require deployment of solutions on complex, powerful computing infrastructure
- Many tools provide a variety of visualization support to show results to end users to support business decision making but are limited to built-in visualization options

- Support domain expert end users to better capture their requirements about target domain problems
- Better support for complex and large datasets, including handling partial and incomplete datasets
- Need both simplicity for non-experts with no data science and programming knowledge, and support for expansion and tailoring for data science experts need to be provided
- Want tool features to capture requirements and changes in requirements as well as adapting the solution based on these changes
- Need scaling and distribution for many real-world applications while balancing this against limited end user knowledge of computing platforms
- Further enhance information visualization capabilities including interactive exploration and end user specification of complex visualizations for the target domain.

Conclusions

- Data analytics phases can be divided to DataOps, AIOps, and DevOps
- Leading data analytics tools address some of these tasks
- Most current tools currently focus on
  - data analytics and machine learning
  - modeling and implementation
  - visualisation
- Many existing tools are complicated for a domain expert with no data science and programming background
- Many are not designed to allow for collaboration between the key stakeholders (team members)