Spring 2020

Big Data for Business Model Renovation, Current Machine Learning Regression Model and Ethical Issue in the Big Data Industry

Chen Chien Wong

MIS 599: Creative Component/Research Paper:

Big Data for Business Model Renovation, Current Machine Learning Regression Model and Ethical Issue in the Big Data Industry

Chen Chien Wong (Jonathan)

Introduction

The global information network generates trillions bytes of data a single day. In just one day, 500 million tweets are posted on twitter, 294 billion email are sent, 4 petabytes of data are created on Facebook and over 5.6 billion google search queries. The constant of web searches, emails, e-commerce transactions, chats, blog posts, social media feeds and data streams from smart devices generated a continuous stream of structure, un-structure, and semi-structure data. These all can be referred as big data. This explosion of data could create challenges for many organization as they struggle to overcome the complexity of information overload. Yet, many organizations overcome the challenges and make use of big data. The defining characteristics of winning companies in this big data age is the ability to capture and analyze the wealth of information available and quickly convert the information to actionable insights.

Big Data insights could improve the patient diagnostic in the hospital, enable retailer to deliver unique customer experience, help bank to detect fraudulent activities before it happens and dynamic improvement of insurance premium by driver's behavior. These are just a few examples of big data implication to our current society. The demand of big data expands the entire spectrum of big data technologies including environments (on premise and cloud), applications, data warehouses, business intelligences, analytics, and integration platforms.

> *"Without big data analytics, companies are blind and deaf, wandering out onto the web like deer on a freeway."*
>
> – By Geoffrey Moore, an American Management Consultant and Author

Big data able to deliver both the company and the customer through a deeper understanding of customer behavior, brand performance and market development. Today, big data offers many opportunities for companies to use analytics in making better decision, and achieving new level of competitive advantages.

In the era of digitalization and unconventional big data analytics, enterprises have begun to derive business value from big data even to the extents of reshaping the business model. This paper will include the brief summary and review from how Lufthansa capitalized big data for business model renovation and embracing customer as value co-creature. Lufthansa has transitioned to a "renovated" airline business model enabled by big data, which we call "Amazon

In the Air" (AIA)—positioning Lufthansa as the Amazon of the airline industry. The study is important as it allow us to understand how big companies discover big data value and embedding it in their business model.

In this paper, a series of technical comparison will be performed to compare the performance of machine learning algorithms in Azure to other traditional statistical analytical tools such as R and Jmp. The General Linear Model (GLM) is a useful framework for comparing how several variables affect different continuous variables. For the purposes of keeping the data constant, the same data set – Airbnb Superhost in Seattle – will be used for all three – Azure Machine Learning, R and Jmp – systems. In term of comparison, three mains factors are considered – The model quality, speed of deployment and the effectiveness.

Big data promises much and permeates many areas of our live, but there are also new problems that come up. How can we learn from useful data while still keeping it safe and private? Perhaps, what are the line we should draw? Today, there is so much data that we have to rely on algorithms to manage it. Can we still trust the algorithms? The topics of ethical in the big data topics are controversial and received a lot of attentions. Big data is often a two edged sword. It can be used either for good or bad that often results difficulty when building regulatory regimes in these areas. For example, in the education sector, big data can help to personalize student instruction and track accountability for performance by school. There is also possibility that the school uses data for admissions discrimination. In addition, big data has been criticized as an intrusion into personal privacy, as potentially discriminator and as distorting the power relationship. People have been raising question as data was collected all the time, they have lost some autonomy. Thus, we are at this point where ethical discussion about what rights do people have on data about themselves. Companies have been reselling consumers' data to the secondary market for Big Data. In the paper, we will examines the ethical issue in the Big Data industry such as negative externality of surveillance and destructive demand together with some suggested solution in creating a sustainable big data industry.

A brief overview on *"How Lufthansa Capitalized on Big Data for Business Model Renovation"*

Lufthansa, headquartered in Cologne, Germany, is the largest airline in Europe in terms of passengers carried. Lufthansa has decided to renovate its business model to differentiate from other premium full service carrier and gain competitive advantage in the highly competitive market.

The new Amazon in the Air (AIA) business model introduced by Lufthansa sees customer as co-creators of values. In the traditional business customers are views as "transaction" – an external entities from the company. The model reframes it as the perspective of customer relationship management and fully transactions the airline to service-dominant logic. The business model uses data from customer to build relationship and trust between the customers. Lufthansa obtains information from customer feedbacks, comments and interactions with customer as the fundamental input for strategy development in faster, simpler and more personalizing the customer experiences. Big Data plays a vital role in developing this renovation.

Big data provides an insight about customer and able to customize the services for reaching out to the targeted individual. Lufthansa uses the data to analyze customers' shopping preference, purchasing history, hobbies, where they live and other information in order to predict what a customer is going to need or want in order to personalized the services. This makes the services more engaging as it meets the customer's specific needs of the service and may significantly influencing the purchasing behavior and experience. In the article, Lufthansa helps customer uses the time according to their own preference during the waiting time instead of hard selling or advertising products. In result, this approach generated higher volume of sales.

Big data predictive analytics also incorporated in the Lufthansa's business model that enable innovative way of handling irregular (IRREG) situation. Lufthansa has developed a big data analytics application to facilitate the IRREG activities such as transferring passenger to a different routes, turnaround processes as well as baggage management. The article uses the example that when an aircraft is delayed, a customer often try to get a car with other passengers to a nearby city. The car share could be facilitated by Lufthansa. In a way, Lufthansa and the customer are co-creating a solution for problems caused by the flight delay. Using the customer data, car services information and even considering the customer response on social media in determining the needs of car arrangement and design the appropriate solutions.

Another disruption that an airline has no control over is bad weather situation. Most of the time bad weather would eventually contributing for delays or cancellation. In the article example, it stated that a huge storm was predicted and all flights were canceled. All passengers have been notified the day before storm hit. The company worked with customer to find the alternate travel options or compensation to satisfy customer needs in such unfortunate event. This renovation business model has indeed helped Lufthansa gain competitive advantage and stand-out in the airline business.

Big Data Paradigm and the associated data pipeline

Before delving into technical aspect, we need to first understand some challenges in big data. Following are some challenges that enterprises need to deal with whenever they are constructing big data workflow:

1. Data is stored in different formats and different location and making it hard to integrate. There are data that are stored on-premises (such as Oracle data warehouse and flat file) and in cloud (such as Amazon s3 and Azure data warehouse). When some of the data in the cloud and some of the data is on premise, it is extremely difficult to move data back and forth and integrate it into one data workflow.

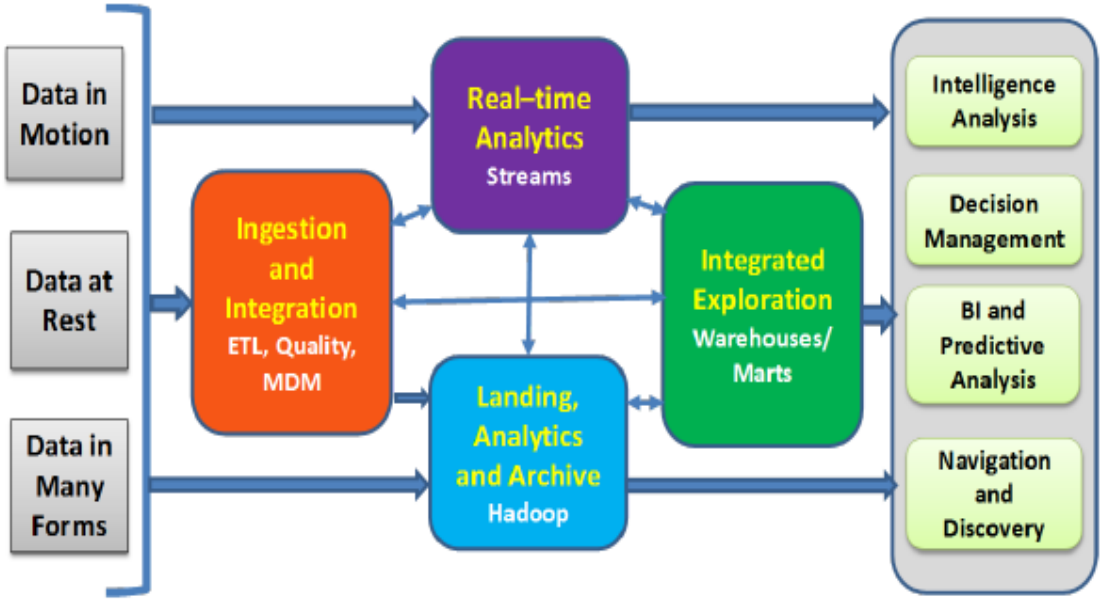
2. Second challenge is that data workflows require complex dependencies. It is difficult to make the pipeline and workflow contingent on all the specified conditions. For example, there may be workflow that requires that completion on step one before proceeding to step two or there may be workflow that need to be run once a month or perhaps at the certain time of the day.

3. Third challenge is that enterprises need to deal with workflow management when things go wrong and must be handle exceptions. Things often doesn't go smoothly. There can be anything that can go wrong such as quality issue in the data or human error. In designing the data workflow, such thought processes need to be considered in the model. Certain questions need to be asked: do you want to retry in the case of an error or an exception? What if a dependent step is taking longer than expected? Do you want to be notified when something goes wrong?

4. There are commercial tool in the market that attempt to do data integration and data analytics. However, many of them are not a good fit for the kind of data and the kind of things that tailored to the company. The tool is very expensive with upfront license fee and many of them are designed to scale up to bigger machines rather than scale out.

Today, company like amazon has invented services called AWS data pipeline that will address many of the challenges. It is usually started with an input data node with attachment of a precondition. The next step in the pipeline is the activity which is something to do with that data. Attach to that activity, there may be some certain failure and delay notification to alert user when something has gone not as expected. Eventually, there will have an output data node that could be a result of predictive analysis or insight that could help management to draw decision. Below is the simplified abstraction of a data pipeline at Lufthansa.

Figure 2: Big Data Paradigm Adoption and Data Flows at Lufthansa



From the left is the data includes data-in-motion, data-at-rest and data in different formats are ingested, cleansed and integrated through a traditional ETL pipeline. Generally, the store operation data in a relational database is used to serve the online transaction. Quality data will move and store in data warehouses in the master data management system either by batches or instantly to do offline analysis. This traditional approaches is somewhat limiting as it takes a longer time to get data analyzed and thus limiting the responsiveness. Previously, Lufthansa relied on small data systems to perform analytics on only internal, structural and transactional

data which was very limited to obtain out of the box insights to gain competitive advantages. Therefore, Lufthansa created this data flow which has secondary data stores and indexes. Raw data like structural and un-structural are archived in a Hadoop system that can handle the large volumes. They can be used at primary data sources. It uses existing data and modifying it in these indexes to perform specialized queries such as text-based queries and graph-based queries. This category result in a lot more destinations for data instead of having primary data stored in one location and maybe ended up in data warehouses. This category allows the data ending in multiple destination data stores along. By this model, the company is now able to combine multiple data source like social media data with their transaction history to understand customer shopping preferences, hobbies, demographical information in order to personalize the experiences. Lufthansa can better forecast and prediction on customer need (e.g., travel to Asia to see their parents or for a business trip) or want (e.g., a vacation in Hawaii) or do (e.g., order a movie on the flight) or buy (e.g., an expensive anniversary gift)

Along with all of this is the trend of real-time analytics. Data needs to be moved between these systems continuously at low latency to perform the result effectively. The shift toward real-time has also brought along with it the rise of stream processing. This means the change in data model that requires all the systems that handle real time data to process it continuously. An integrated exploration uses data from stream events and combined the traditional data from data warehouses and Hadoop to perform analytics. This also implies that the data pipeline need to provide a steady stream of events. In the diagram, the model shows that the results from four modules (real-time analytics, landing, analytics and archive, and integrated exploration) are then displayed in the helping management to make decision, intelligence analysis, business intelligence and predictive modeling, and navigation and discovery modules. The employee or manager of Lufthansa can then interact with these modules and yield insights. For instance, a flight plan with compliance of all safety rule and regulations requirement is generated for each flight. Lufthansa used historical plans with actual performance and mining the data of flight histories to yield insights for minimizing flight duration, improving aircraft efficiency and minimizing maintenance duration. As a result, this reduced the overall operational costs such as fuel. Ideally, optimal flight plans will be automatically generated and updated based on the environment at both departure time and during the flight.

Statistical Tool Comparison against Machine Learning - Regression

After discussing cases on how big data can be influencing the existing business structure, let's put this into a more practical sense. Big data flows as we discussed earlier include storage, ingestion and extraction tools. Whereas machine learning provides computer ability to learning without being explicitly programmed. In another words, machines learning is teaching a machine to intake unknown input and give desirable outputs by using machine learning algorithms. Both big data and machine learning could come hand in hand. Many data inflow every second and humans certainly not able to analyze the data every minutes we obtain new data. Thus, through machine learning with an appropriate set up of inflowing the data, it can be set up to automatically look for specific types of data and parameters and their relationship between them big data can't see the relationship between existing pieces of data with the same depth that machine learning can. In term of effectiveness, it is definitely a win.

Machine learning with the combination of good data is often used in the field of artificial intelligence by using software application to train continuously and increase the accuracy for the expecting outcome. In layman's terms, Machine Learning is the way to educating computers on how to perform complex tasks that humans don't know how to accomplish. Machine Learning field is so vast and popular these days that there are a lot of machine learning activities appending in our daily life and soon it will become an integral part of our daily routine.

In this part of the paper, a linear regression model - linear relationship between one or more independent variables – will be created by machine learning software using Microsoft azure. In additional, two other statistical software, R and Jmp will be used to create a linear regression model manually. For the purposes of keeping the data constant across three software, the same data set – Airbnb Superhost in Seattle – will be used. In term of comparison, three mains factors are considered –model quality, speed of deployment and the effectiveness. Linear regression is one of the most commonly used statistical method in the real work when dealing with structural data. It has been adopted in machine learning and enhanced with many new method for fitting the line and measuring error. Azure machine learning supports a variety of regression model in addition to linear regression and it is designed in a way where user does not need to extensive coding to have it work.

Seattle Airbnb Listing Data

The data will be used for the experiment is from Inside Airbnb which is an independent, non-commercial set of tools and data that allows you to explore how Airbnb is really being used in cities around the world. Seattle listing Inside Airbnb is selected due to the volume of the records and being one of the location that has multiple residential housing type. Many fundamental questions can be asked with the listing data about Airbnb in any neighborhood or across the city as a whole. Questions such as: "How many listings are in my neighborhood", "How many houses and apartments are being rented out frequently to tourists and not to long-term residents?", "How much are hosts making from renting to tourists (compare that to long-term rentals)?", "Which hosts are running a business with multiple listings?". As standing from a data analyst point of view in helping the host to boost the profit, we are going to perform a series of analysis to determine what factors are more important to the customer. One secret that every Airbnb host know, is to earn a Superhost badge.

The Superhost program celebrates and rewards Airbnb's top-rated and most experienced hosts. In return, the host would earn extra money, attract more guests and access to exclusive rewards. Superhosts are generally more visible and trustable from guest which can mean more earning.

The data behind the Inside Airbnb site is sourced from publicly available information from the Airbnb site. The data has been analyzed, cleansed and aggregated where appropriate to facilitate public discussion. Logistic regression will be performed using the data available to explain and predict the factors that affect the chances of obtaining a superhost badge. The data contains 9040 rows and 106 columns. Following are some important attributes with explanation:

Target Variable: SuperHost (Y/N) – superhost status

Predictors:

idlisting ID

namename of the listing

host_idhost ID

host_namename of the host

neighbourhood_grouplocation

neighbourhoodarea

latitudelatitude coordinates

longitudelongitude coordinates

room_typelisting space type

priceprice in dollars

minimum_nightsamount of nights minimum

number_of_reviewsnumber of reviews

last_reviewlatest review

reviews_per_monthnumber of reviews per month

calculated_host_listings_countamount of listing per host

availability_365number of days when listing is available for booking

Overview of tools

R studio, Jmp and Microsoft Azure Machine Learning are statistical tool used by analytics pros, statisticians, and data scientists in many organization. There are many difference from one product over the other. First, let's start by defining what each product has to offer:

- Microsoft Azure Machine Learning is a relatively new service offered by Microsoft on their Azure cloud platform. Azure Machine Learning is also GUI-based and is used for constructing and operationalizing machine learning experiments on Azure. This is hosted in the cloud, connected to your Azure account.
- R studio is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.
- JMP is a software program used for statistical analysis. It is created by SAS Institute Inc.

Cost and setup

	R Studio	Jmp	Microsoft Azure ML
Pricing	RStudio is a free, open source IDE (integrated development environment) for R	JMP Statistical Software pricing starts at $1785.00 per year, per user.	The cost to use Azure Machine Learning is dependent on how much you choose to use it.There are two price tiers for the service: Free and Standard. For the Standard tier, you pay by the number of users that need access to the service and by the number of hours each user uses the tool. This cost is still very low, ranging from around **$9.99** per month for minimal use to the extreme of

			$**729.99** per month if the user were to run experiments 24/7 for the entire month.
Platform	Installed - Mac Installed - Windows	Installed - Mac Installed - Windows	*Cloud, SaaS, Web Installed - Windows
Training	Documentation	In Person, Live Online, Webinars, Documentation	Live Online, Documentation

*There is no need to install anything in order to use Azure Machine Learning Studio, as everything is in the cloud. Therefore, users can work even when on the go, making the platform convenient mobile workers.

Sourcing data

	R Studio	Jmp	Microsoft Azure ML
Source	CSV, TXT, HTML, and Other Common Files into R (JSON, XML) SPSS, Stata, SaS file Databases and Other Sources (webscrapping) with the help of package	text, .csv, .jsl, .dat, .tsv, and .jrn	Azure SQL Database or On-Premises SQL Database Azure Blob Storage, Azure Document DB, or Azure Table Data Feed Provider (OData) Web URL via HTTP Hive Query Manually-Entered Data (copy and paste from a .csv, .tsv, .arff, etc.)

Coding

	R Studio	Jmp	Microsoft Azure ML
Coding	Required	No	No, can be coded

Drag and Drop	No	Yes	Yes

In the Mircosoft Azure ML workspace, users can choose to code their R or Python scripts. It can be fully integrated with the rest of the experiment's workflow. For R and Python, many common packages are already installed and can be referenced from within your custom script. For packages not installed, you can always upload the .zip file of the package and reference the package in an alternative way in the R script.

Time consumption and performance

In this part, linear regression prediction models are performed in all three programs with the Seattle Airbnb dataset. The purpose of the experiment is to compare the time consumption that each software will take to come out with the best model. Four major sections are compared – data preparation, model creation, model evaluation and lastly deployment.

(Mins)	R Studio	Jmp	Microsoft Azure ML
Data Preparation			
Cleaning	30	30	30
Dimension reduction	240	240	
Model Creation			
GLM models (including experimental cases)	40	25	
Model Evaluation			
Evaluation and Selection	30	25	30
Model Deployment	0	0	5
Total (mins)	340	320	65

Pre-processing and cleaning data are important tasks that must be conducted before a dataset can be used for model training. Raw data is often noisy and unreliable, and may be missing values. The data has been cleansed and aggregated by Airbnb where appropriate to facilitate public discussion. This essentially reduced the time being used for data cleansing as it contains way lesser irregularities or corrupt data. All three software is recorded the same used time of 30mins.

Dimension reduction recorded as 240mins in R and Jmp are conducted in WEKA software. Feature selection is divided into attribute evaluator and search method. Multiple iterations are run for each combination which has multiple techniques from which to choose. The attribute evaluator is the method by which each column in Airbnb Seattle dataset is evaluated in the context of the output variable. The search method is the method by which to narrow down different combinations of column in the dataset. The result from each iterations are summarized and documented in the appendix.

R takes 70 minutes for the remaining glm model creation and model evaluation. Jmp take 50 minutes. This result in total of 340 minutes for R and 320 minutes for Jmp. While the entire process for Microsoft Azure ML only take 65 minutes which is approximately 5 times faster. The setup of ML are snipped and documented in the appendix.

In Azure ML, the process iterates through different machine learning algorithms and hyper parameter settings, adhering to the predefined constraints. It chooses the best-fit model by optimizing an accuracy metric. In the sample run, 53 iterations are run in less than 30 minutes. Each iteration has a widget that allows to monitor the performance of the table visually in graph along with training accuracy metrics and metadata. Model 37 is the best model out of all the iteration statistically. The output in the dashboard returns and suggest the best run and the fitted model for the last for invocation. Lastly, the estimated time to finish deployment is about five minutes to configure the image and deploy. Again, the steps are being documented in the appendix.

Ethics Issue in Big Data Industry

The algorithms brings us to our first concern with Big Data which is biasness. It is too big for a lot of the usual traditional programs to process. It is too large that often managed by algorithms which inadvertently introduce bias. Algorithms could handle and process data but with higher stakes. Ones used to determine mortgage and insurance rate, or assess the risk someone will do something illegal in the future which might pick up on things like race or other minority statuses. This is real and not uncommon around us. Judges in the United States use risk assessment programs while making sentencing decisions. A commonly used program is called COMPAS, which was created by the company Equivant. It basically gives a score of how likely a person is to commit another crime within two years. In 2016, ProPublica published an investigation on COMPAS. They inspected at the score of 7,000 people who has been arrested in Broward County, Florida. The score were compared with whether those people actually ended up committing crimes again within two years. ProPublica found that

> *"The formula was particularly likely to falsely flag black defendants as future criminal, wrongly labeling them this way at almost twice the rate as white defendants. White Defendants were mislabeled as low risk more often than black defendants."*

The company, Equivant then claimed that in order to make scores as accurate as possible, certain factor had to be included that could correlate with race, ethnicity or demographical background. The data is neutral but the correlation and result may be biases and skewed toward a certain group of people. Data that uses to create our algorithms need to be as representative as possible. One of the many suggestion to create this type of unbiased algorithms aren't always practical. If we wanted to build an algorithm that predict the success of executive members, and data provided to train the model only gave the examples of male who succeeded. Algorithms will have a bias. Many suggests that the model need to supply with good and unbiased data – Males who succeeded and failed as well as Females. However, it can be hard to determine what exactly an algorithms is doing from the raw data to the output. If the outcome is predetermined, then there is no need of analyzing and creating the predictive model. Isn't that contradictive to primary purpose of creating the algorithms if we already known that such predictor could yield a "bias" or "unbias" target outcome?

Another danger emerging from Big Data are the security, as all the data out there means there are a lot of information that can be stolen. Better technology does allow for more protection like encryption but it also exposes our data to wider scale breaches. Hacker may after customer personal information that can be used to set up lines of credits like when Equifax was hacked in 2017.

The next issue is transparency. Data is traded between companies in the secondary market with the intent of reaching out a broader customer base. The trading trend raised when there are the need of consumer data – consumer facing companies are being pressured to collect and sell increasing amount of data collected with lower standard. This is due to the lack of internal security control and law in places that allows unauthorized personal data handle customer data and even extracting it to make profit out of it. As these questionable norms and practices within the industry become more common, it becomes a destructive demand. There are companies that focusing on reselling of consumer data. The website application can become a bait for big data such as flashlight application tracks your location. Such phenomena is happening because the secondary market is more lucrative than the primary. Selling consumer information may earned more or at minimum equal to the profit generated from the primary market. Secondary, firm in the primary market have limited accountability to consumer for their transaction in the secondary market. Most of the time the activity in the secondary market is not visible. Consumer facing organization are currently held no accountable for selling access to consumer data even by market force and their activity in the secondary market is invisible to the primary consumer market. Consumers are simply treated as a mere means to supply the secondary market of information trader. As per Immanuel Kant's categorical imperative theory, the formula of humanity, "Act so that you treat humanity, whether in your own person or in that of another, always as an end and never as a mere means." To use something as a mere means is to use it only for your own benefit with no thought to the interest or benefit of the things that are being used which can be referred to customer in this case. Based on the thought, Kant would suggests that trading consumers' data in the secondary market without consent are not ethical because we are all end-in-ourselves.

Privacy is another concerned of big data. There are all kinds of personal data about an individual that might not want to be shared such as entertainment choice, shopping preference,

web search history, activities even like basic information of location, step counts, heart rate are tracked and recorded. There are a lot of question when thinking about privacy: Who has access to all that information? Who owns the data? Who are they sharing with? What are they doing with it? In 2018, the European Union implemented a law the General Data Protection Regulation (GDPR) that addresses a lot of privacy concern people have with the use of big data. For example, Facebook recently was unable to comply with the stricter EU regulations because of a lack of adequate consent and control for users: Facebook users have no true opt-out mechanism, no valid consent for the transfer of data to third parties and a general lack of control over their data. GDPR requires companies that deal with Big Data to be more transparent about what they are collecting and who can see it. Privacy laws have been around for a long time all over the world, but as pertain to big data, a lot of new consideration need to be re-think and policy maker are still figuring out.

Conclusion and Recommendation

The importance of ethical considerations in big data continues to grow due to the ever increasing uses into all areas of personal and public life in industrialized societies. Earlier, we see that how fast a simple machine learning model in Azure can replace the tradition data scientist work in faster and more efficient manner. There may be some improvement that still need to be improved from time to time increase the human sense to the program. It does open up the new market where big data can reveal unforeseen about our lives and create opportunity for the company to capture the market. Historically, the company only uses transaction history to do research but now with help of big data, company increasingly repurpose other type of data that themselves seems to have nothing to do with health. When combining with other sources, it may ended up reveal something. The potential invasiveness of those new ways of looking at data is one of the central motivating factors. This causes in the space of ethics and also in any approaches of regulating data such as data collection, data analysis, data protection, and privacy law.

Data sharing between universities, hospitals, and other organization share data can be benefit to the society. A health organization survey on risk behaviors, like drug use, could have incredibly valuable results to researches and policy makers. We can try to make it so that data can't be easily connected to the specific person. The obvious first step is to not include people's name, or other unique, identifying personal information. If someone has a rare disease, simply knowing the city where they live might be enough to pin point to figure out who they are. One option to combat this issue is to make sure that there are at least 2 or more subjects that have the same characteristics. This is called k-anonymity. K refers to the number of subjects who share the exact same characteristics. If there are two people with that disease from that city, we have 2-anymity because there were two subjects or records with the same characteristics.

Don't let our excitement about the big data to outpace our caution. Practical solutions to creating mutually beneficial and sustainable relationships within the industry include

(1) Visible data stewardship practices
(2) Greater data due process internally
(3) Using the services of data integrity professional

Reference and Citation

1. Chen, H.-M., Schütz, R., Kazman, R., & Matthes, F. (2017). How Lufthansa Capitalized on Big Data for Business Model Renovation. MIS Quarterly Executive, 1–16.

2. Desjardins, J. (2019, April 17). How much data is generated each day? Retrieved November 5, 2019, from https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/.

3. Larson, J., Angwin, J., Kirchner, L., & Mattu, S. (2019, March 9). How We Analyzed the COMPAS Recidivism Algorithm. Retrieved November 1, 2019, from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

4. Martin, K. E. (2015). Ethical Issues in the Big Data Industry. MIS Quarterly Executive, 1–19.

5. Price, E. (2019, August 12). The EU Might Fine Facebook Billions For GDPR Violations. Retrieved November 1, 2019, from https://www.digitaltrends.com/social-media/facebook-gdpr-decision/.

6. Stahl, B. C., Mittelstadt, B., & Timmermans, J. (2016, February 1). The Ethics of Computing: A Survey of the Computing-Oriented Literature. Retrieved October 31, 2019, from https://www.academia.edu/22379063/The_Ethics_of_Computing_A_Survey_of_the_Computing-Oriented_Literature.

7. SWEENEY, L. A. T. A. N. Y. A. (2002, May 22). a model for protecting privacy. Retrieved November 5, 2019, from https://dataprivacylab.org/dataprivacy/projects/kanonymity/index.html.

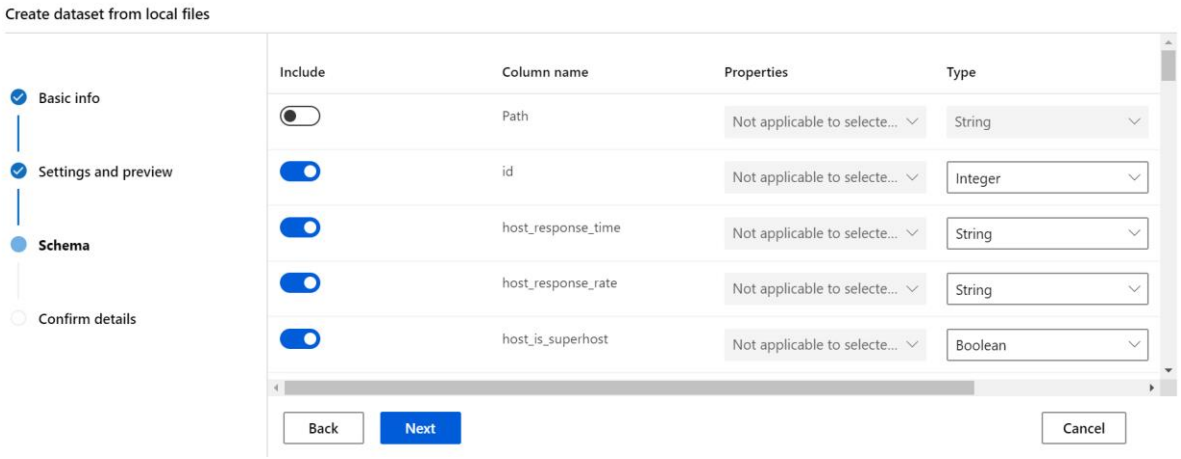
8. Williams, K. (2018, November 20). This R Data Import Tutorial Is Everything You Need. Retrieved January 20, 2020, from https://www.datacamp.com/community/tutorials/r-data-import-tutorial

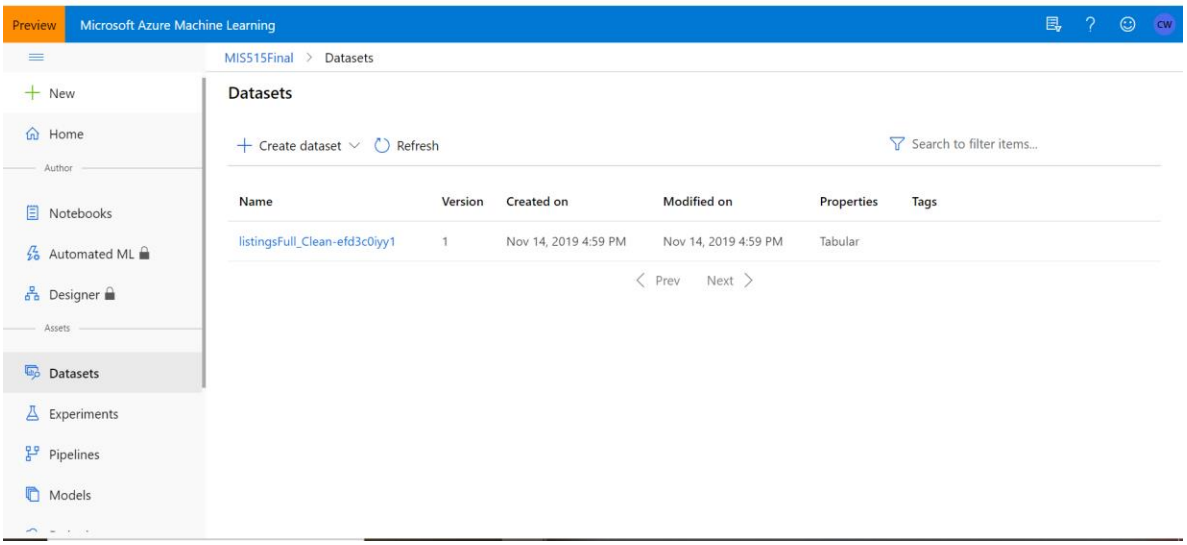
9. Szeliga, M., & Szeliga, M. (2015, January 25). Pros and Cons of Azure Machine Learning. Retrieved from sqlexpert.pl/2015/01/25/pros-and-cons-of-azure-machine-learning/

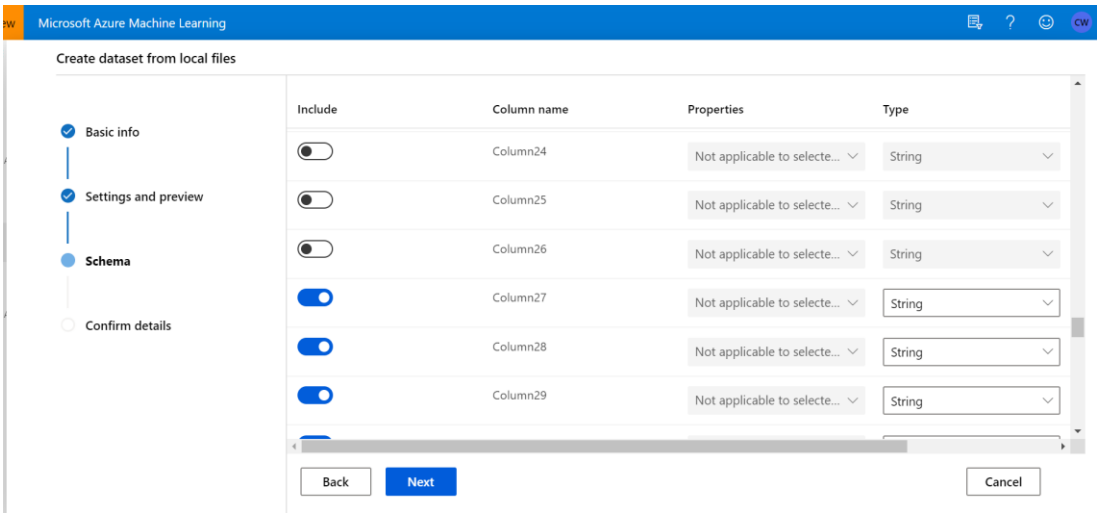
10. Ford, C. (2017, March 2). SAS Enterprise Guide vs. Microsoft Azure Machine Learning. Retrieved March 27, 2020, from https://www.blue-granite.com/blog/sas-enterprise-guide-vs.-microsoft-azure-machine-learning

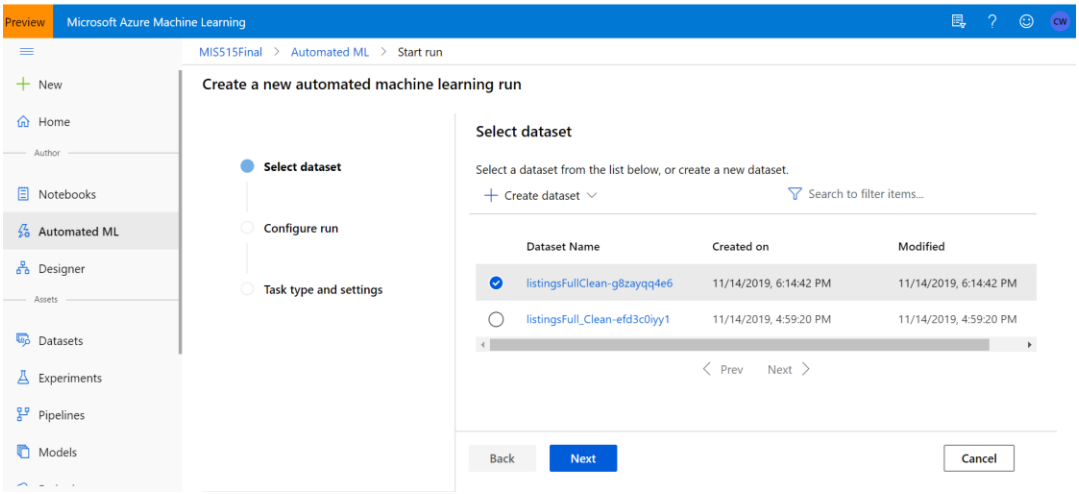
Appendix

ML Azure

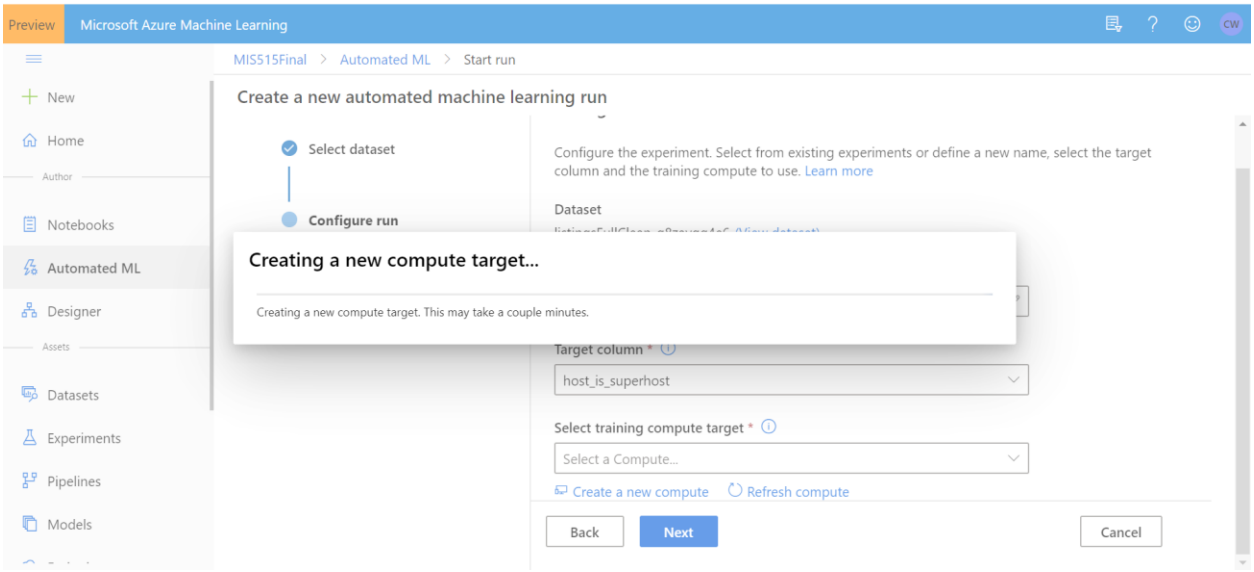




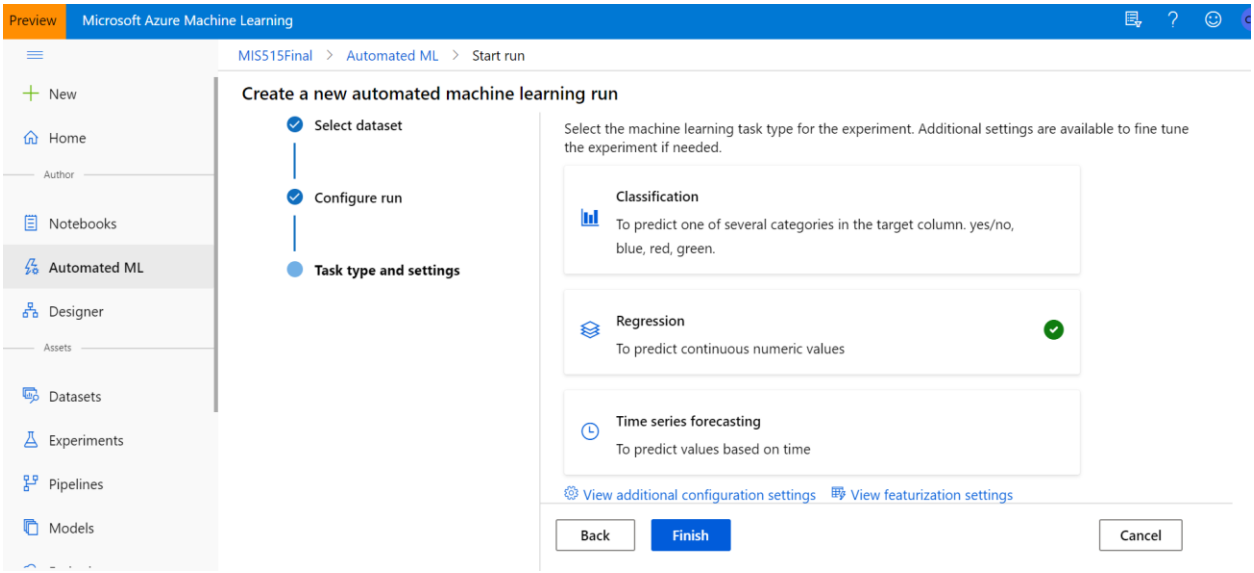




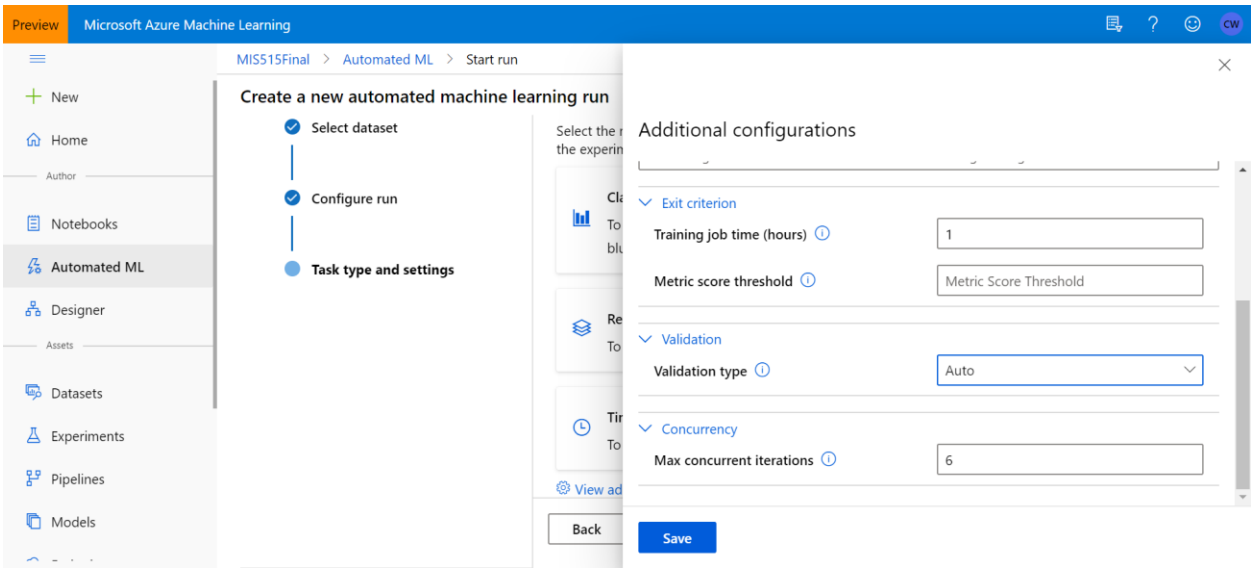
Creating compute target



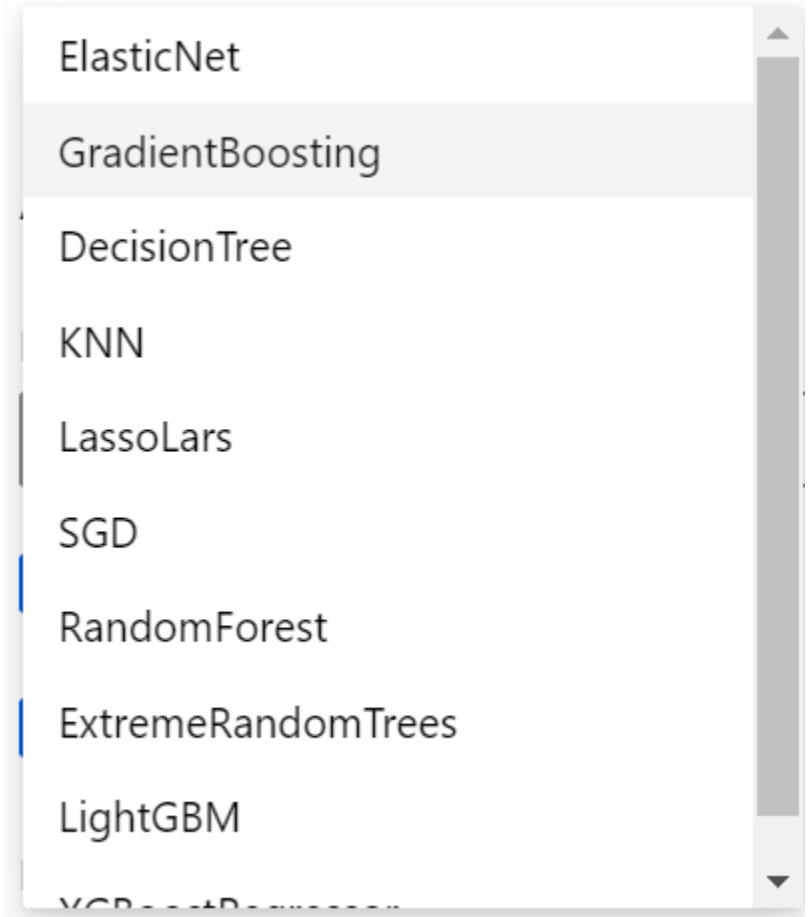
Run regression

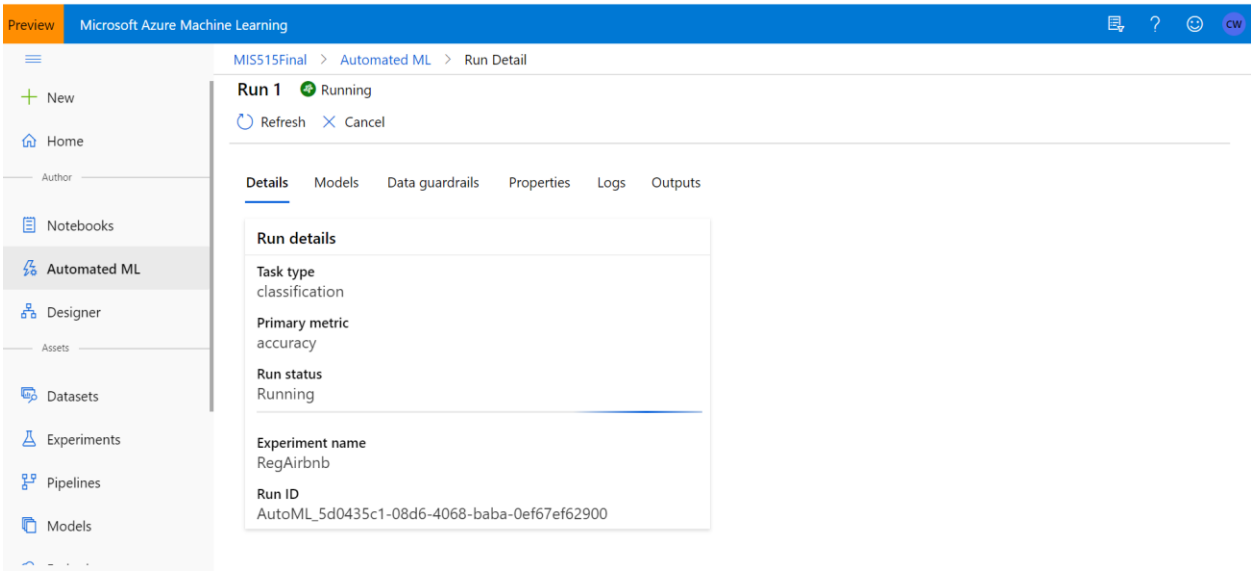


spearman_correlation

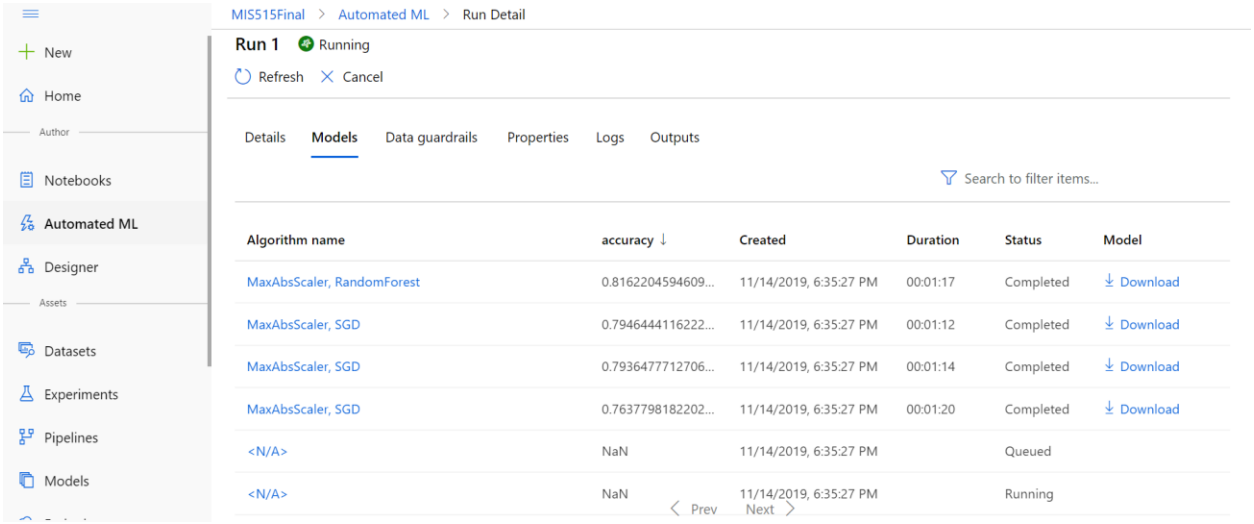


Run all model

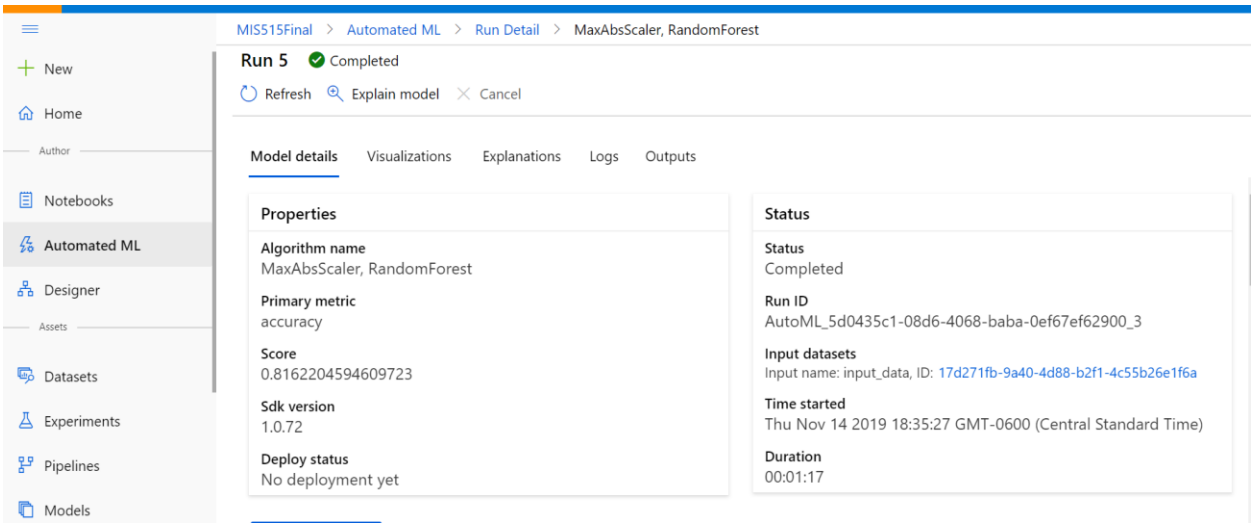




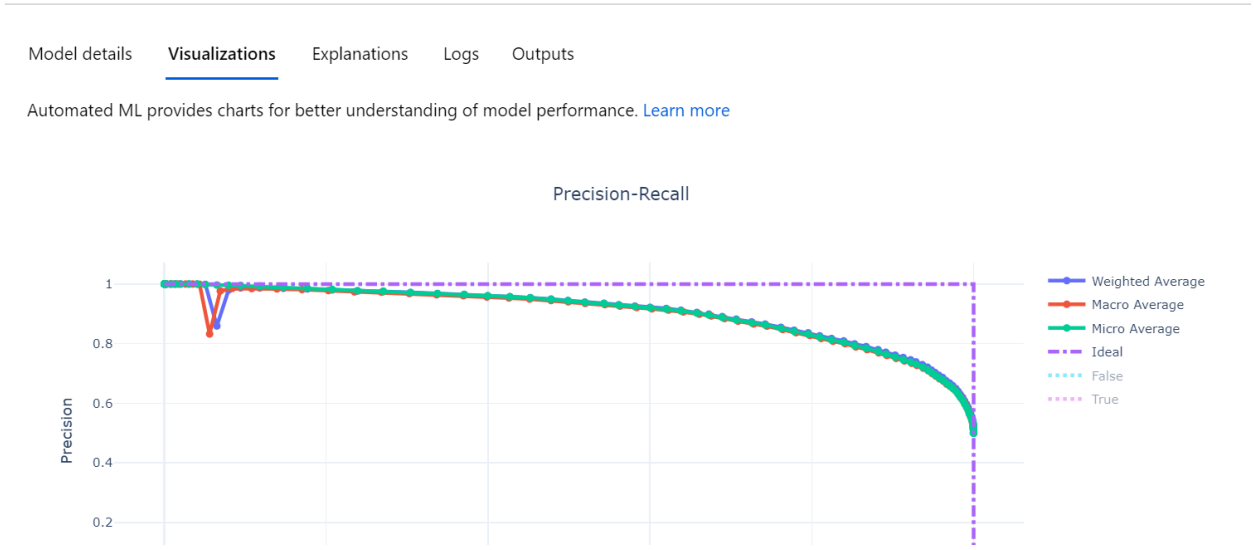
Running model

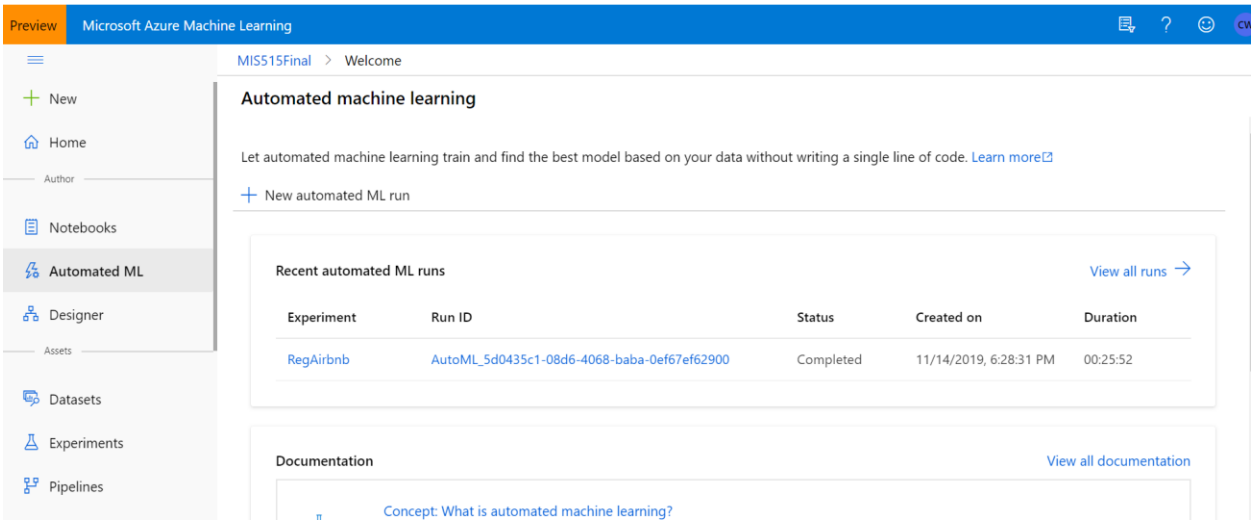


Example RUN5 :

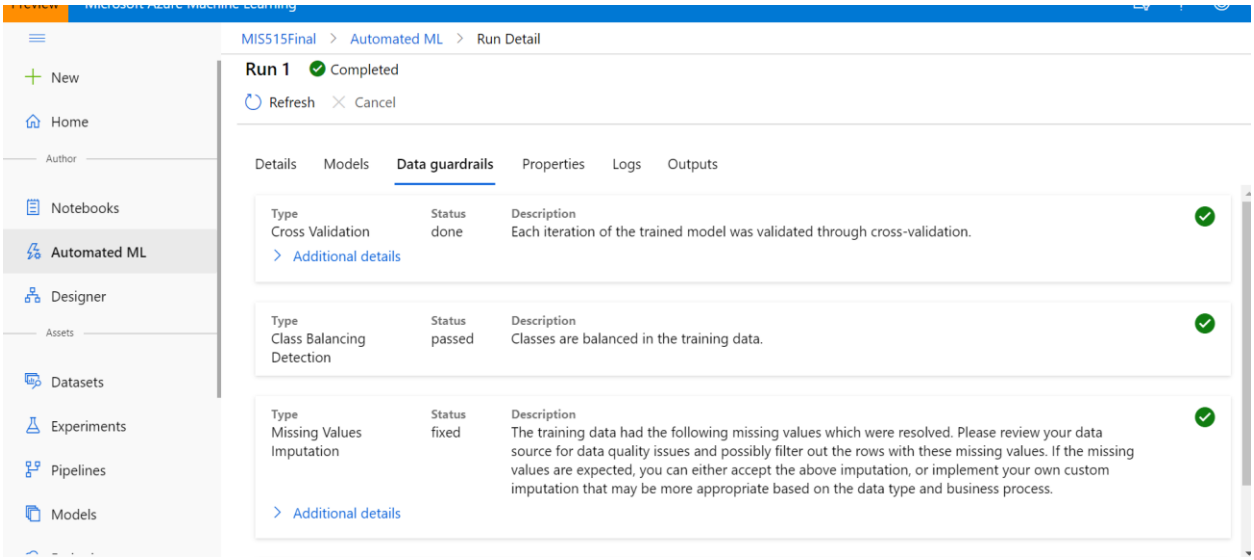


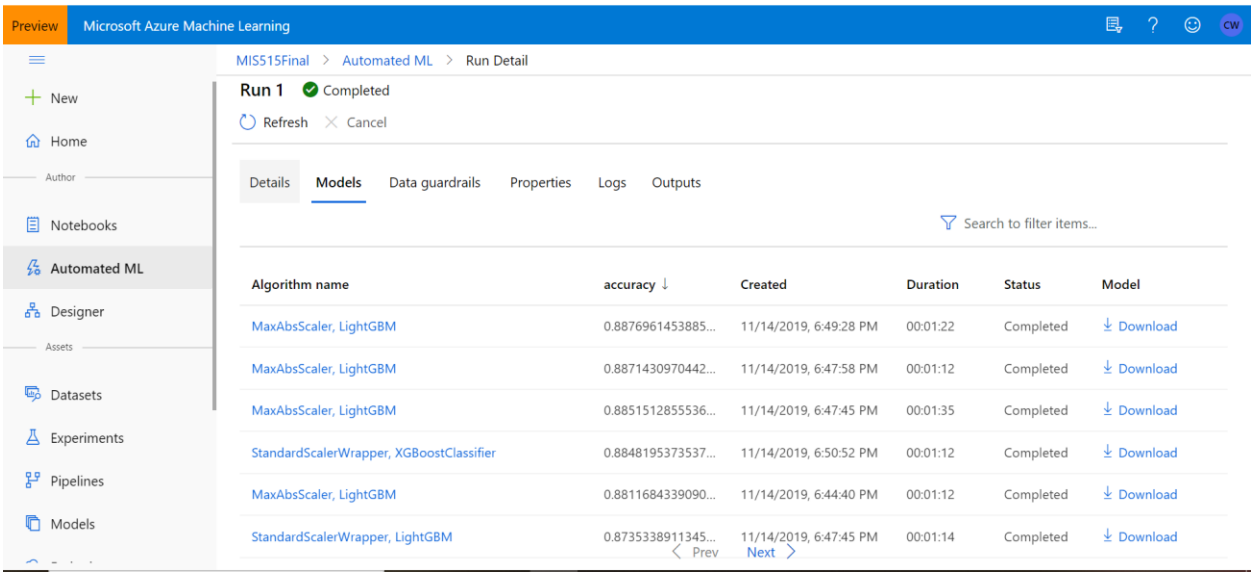
Visualization of the model

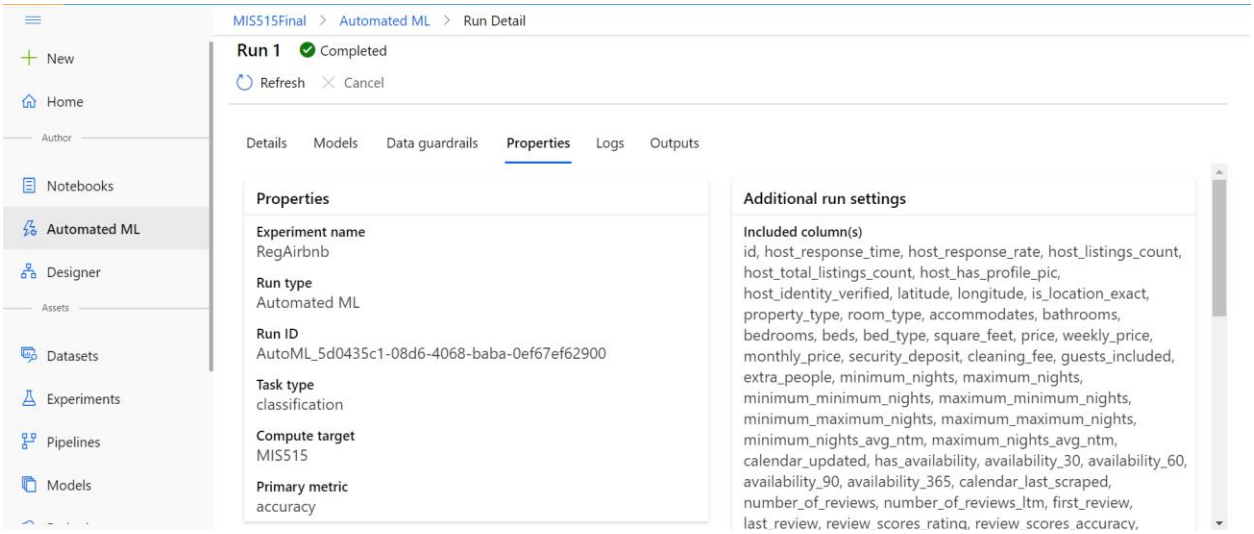




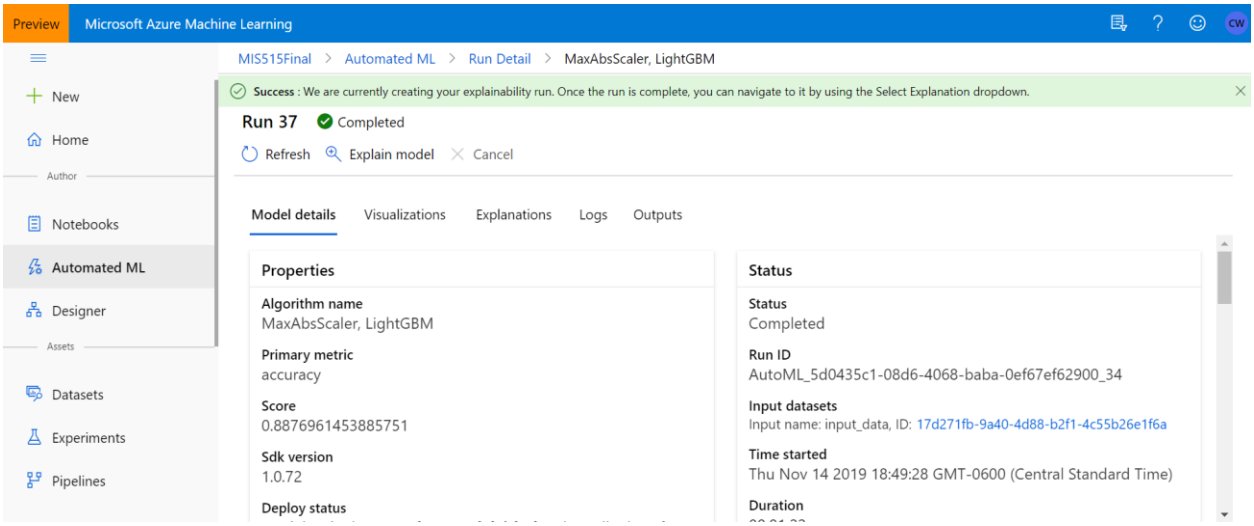
Cleaning







Model 37



Run Metrics

norm_macro_recall0.77025

balanced_accuracy0.88512

AUC_weighted0.95877

average_precision_score_macro0.94296

AUC_micro0.95877

precision_score_weighted0.88788

f1_score_macro0.88477

average_precision_score_micro0.94296

average_precision_score_weighted0.94296

f1_score_micro0.88770

f1_score_weighted0.88774

log_loss0.26068

precision_score_macro0.88452

weighted_accuracy0.89014

recall_score_macro0.88512

accuracy0.88770

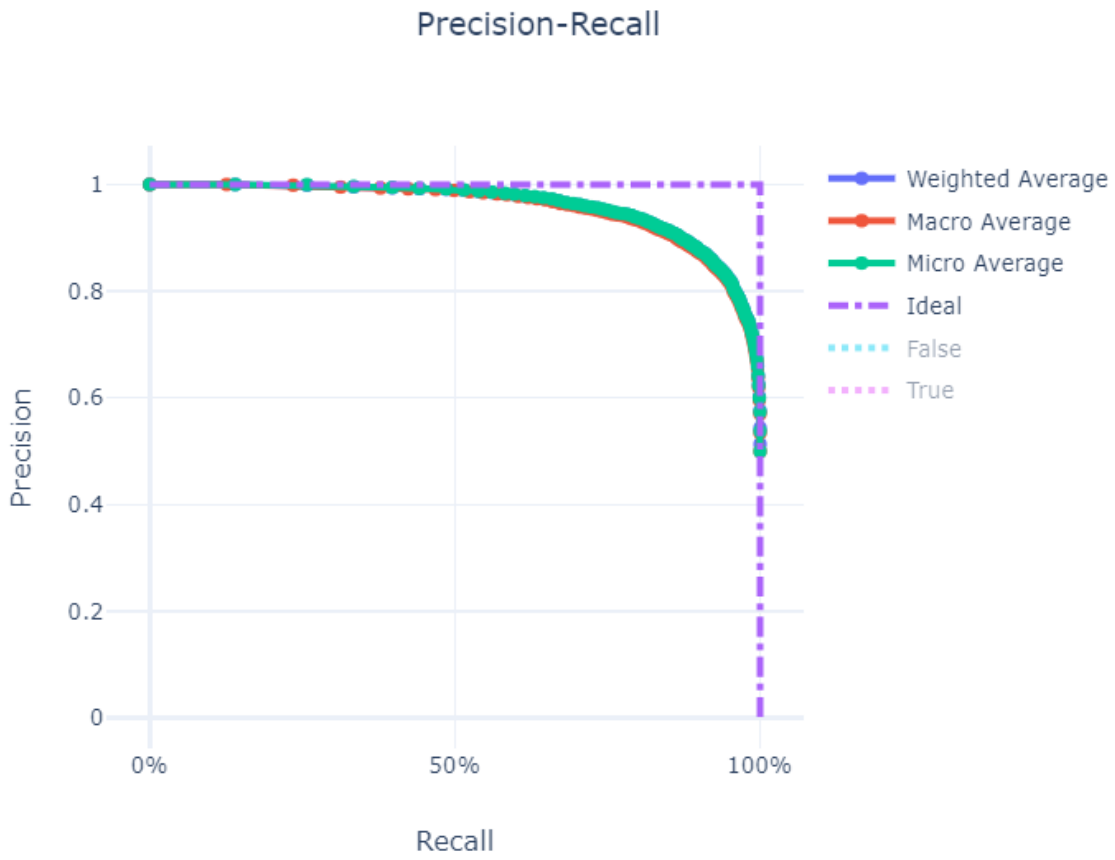
recall_score_micro0.88770

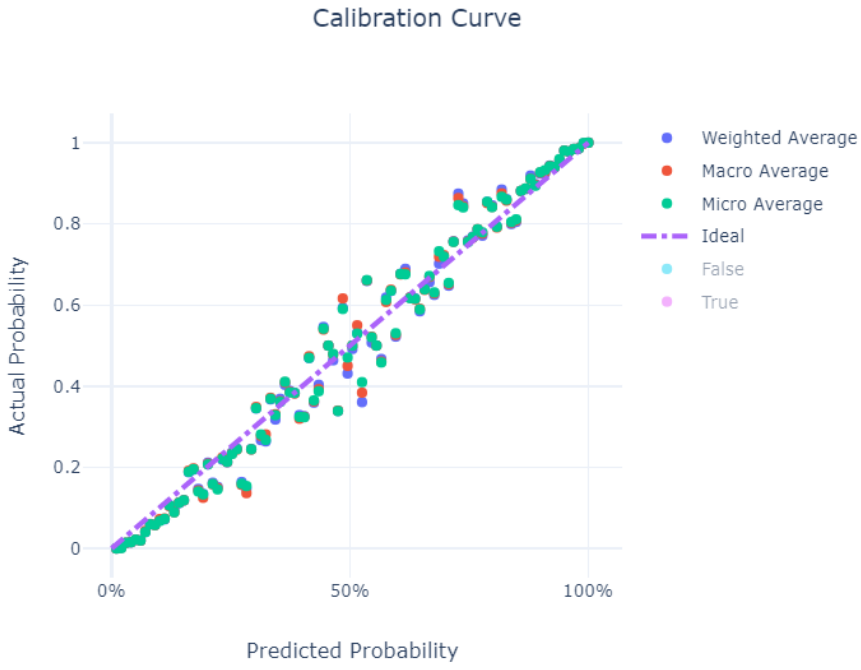
AUC_macro0.95877

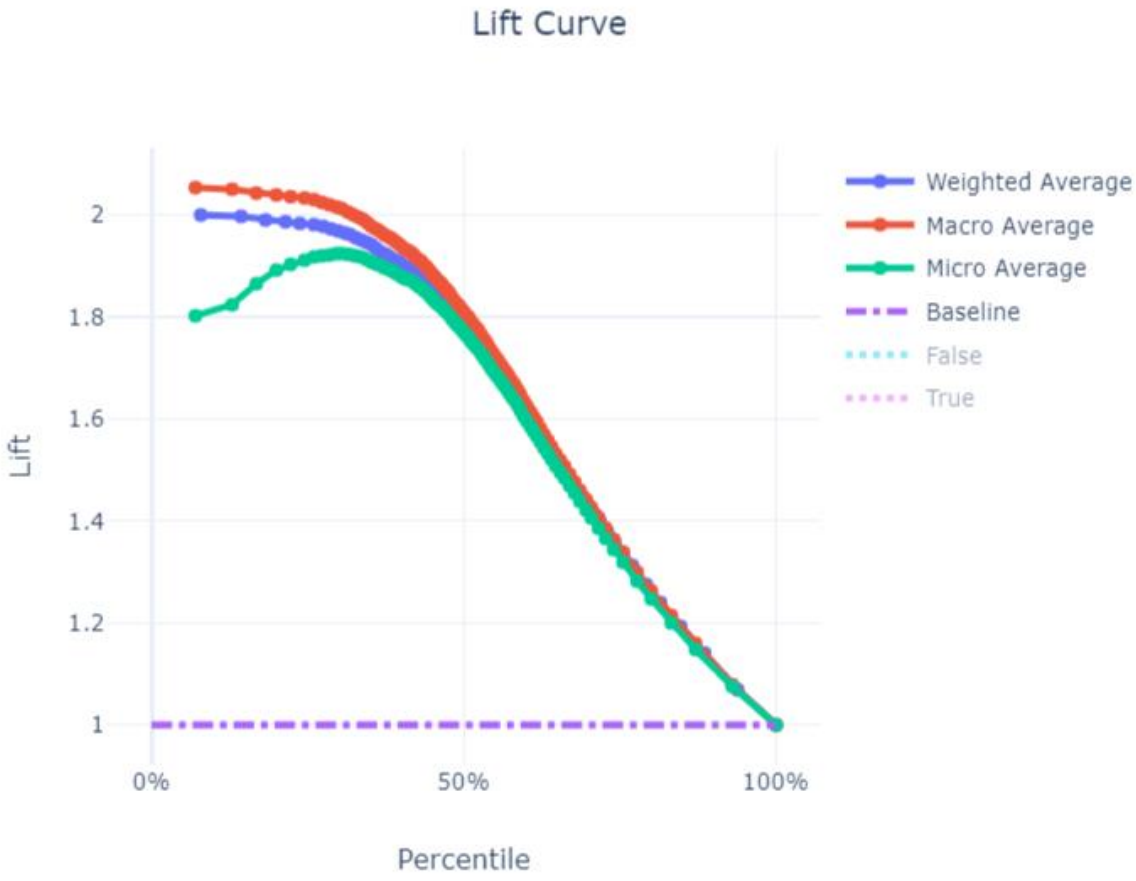
recall_score_weighted0.88770

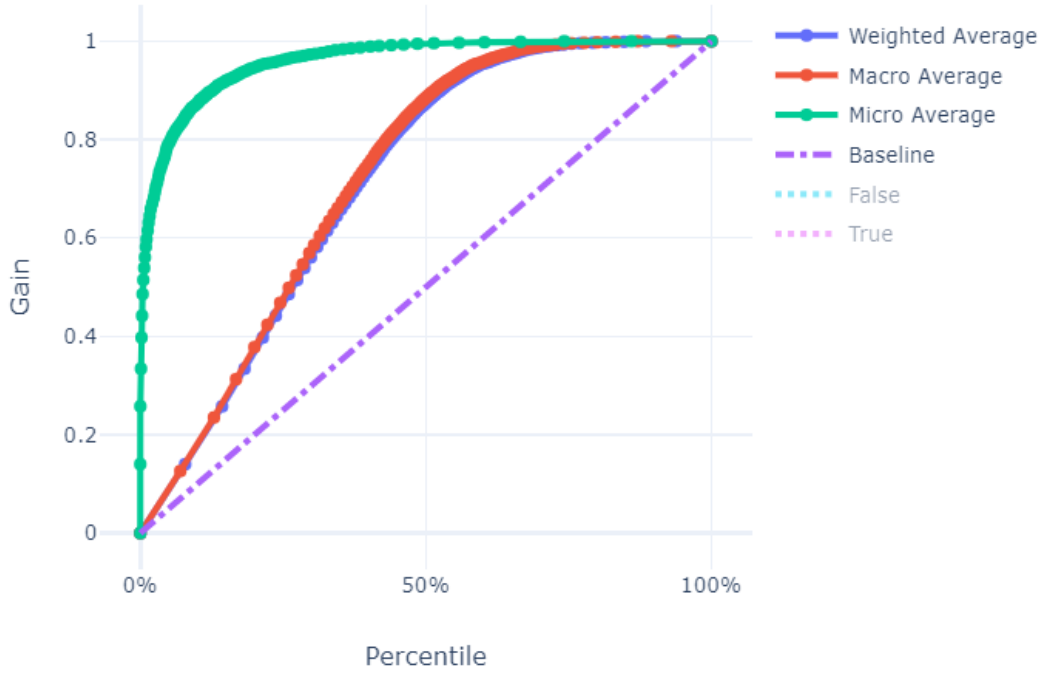
precision_score_micro0.88770

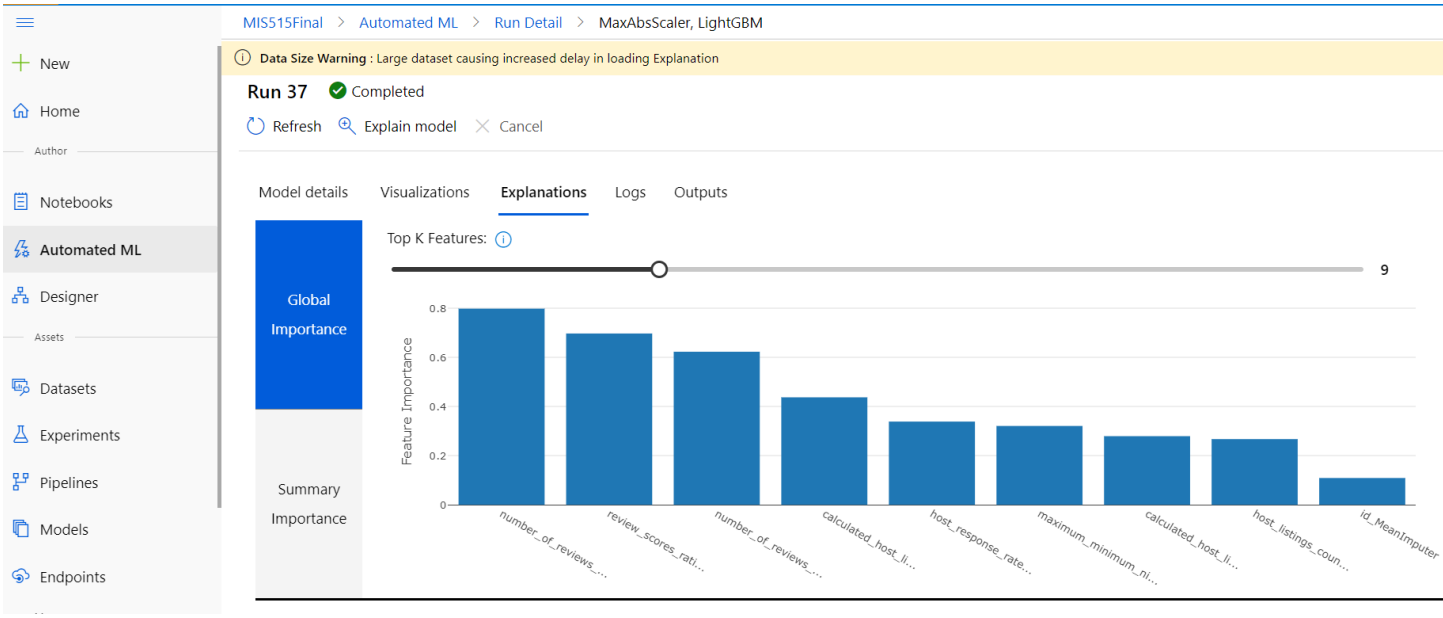
Evaluation of the model











Number of review – 0.798

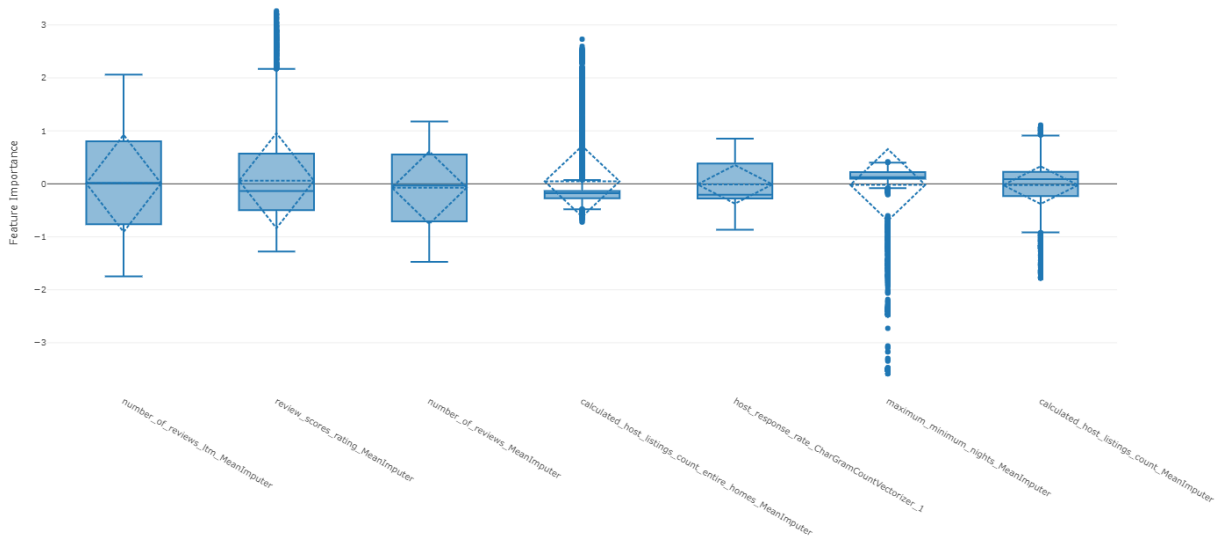
Review rating -0.697

Number of review_ltmputer -0.623

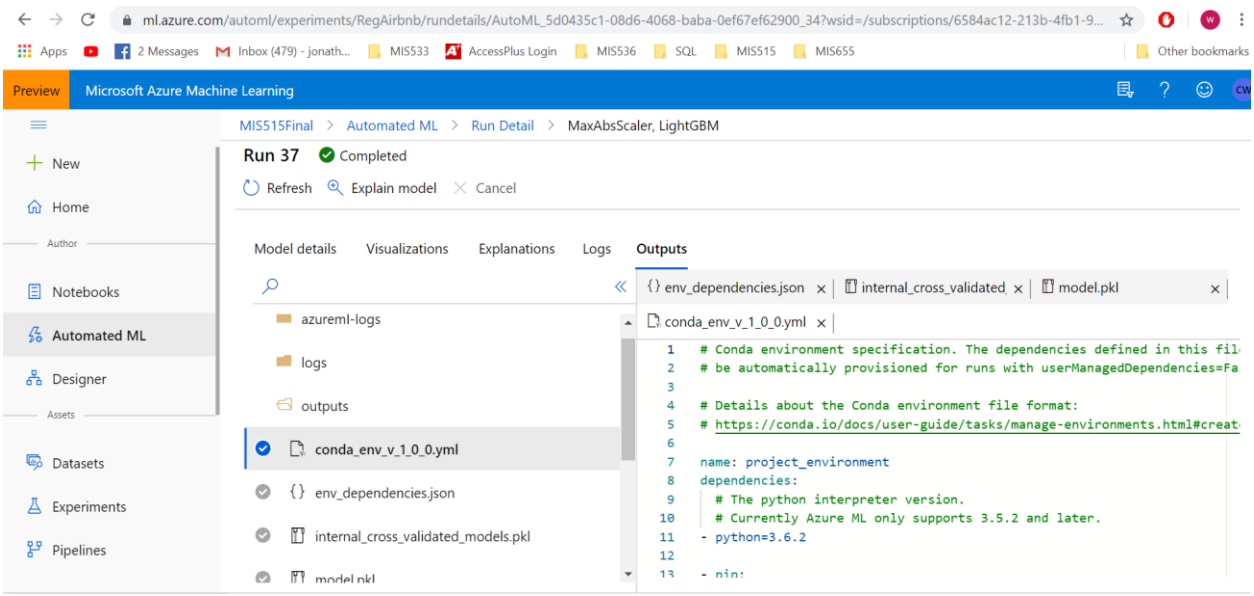
Host listing count_ entire home -0.437

Host response rate – 0.399

Maximum night – 0.321



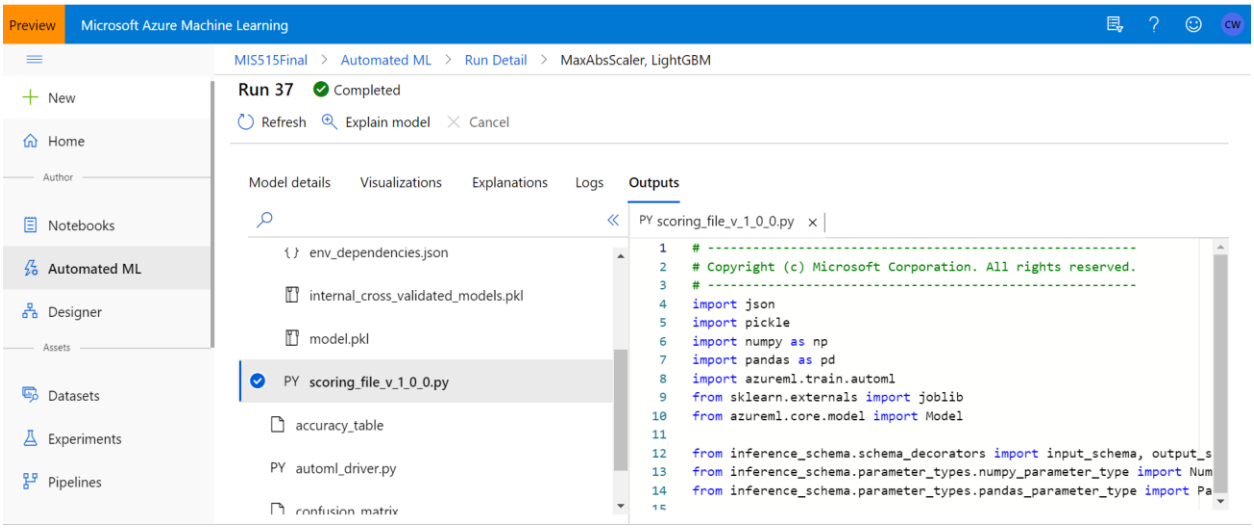
Download the conda script



Conda environment specification. The dependencies defined in this file will

```
# be automatically provisioned for runs with userManagedDependencies=False.

# Details about the Conda environment file format:
# https://conda.io/docs/user-guide/tasks/manage-environments.html#create-env-file-manually

name: project_environment
dependencies:
  # The python interpreter version.
  # Currently Azure ML only supports 3.5.2 and later.
- python=3.6.2

- pip:
  - azureml-train-automl==1.0.72
  - inference-schema
  - azureml-explain-model==1.0.72
  - azureml-defaults==1.0.72
- numpy>=1.16.0,<=1.16.2
- pandas
- scikit-learn
- py-xgboost<=0.80
- fbprophet==0.5
- psutil>=5.2.2,<6.0.0
channels:
- conda-forge
```

Download script

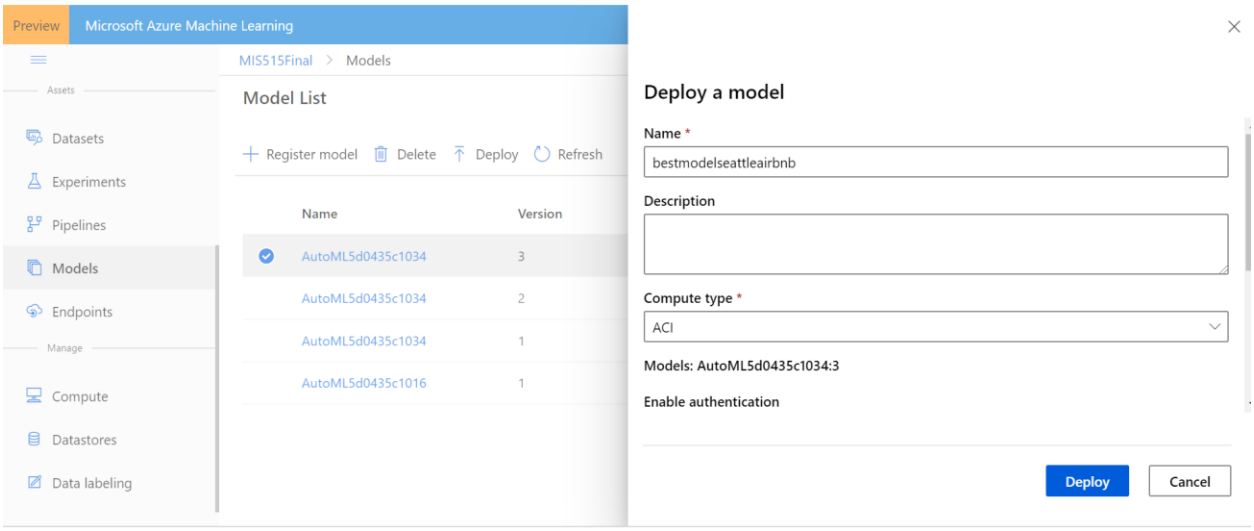



```
# ---------------------------------------------------------
# Copyright (c) Microsoft Corporation. All rights reserved.
# ---------------------------------------------------------
import json
import pickle
import numpy as np
import pandas as pd
import azureml.train.automl
from sklearn.externals import joblib
from azureml.core.model import Model

from inference_schema.schema_decorators import input_schema, output_schema
from inference_schema.parameter_types.numpy_parameter_type import NumpyParameterType
from inference_schema.parameter_types.pandas_parameter_type import PandasParameterType


input_sample = pd.DataFrame(data=[{'id': 2318.0, 'host_response_time': 'within an hour',
'host_response_rate': '1', 'host_listings_count': 2.0, 'host_total_listings_count': 2.0,
'host_has_profile_pic': True, 'host_identity_verified': False, 'latitude': 47.61082, 'longitude': -
122.29082, 'is_location_exact': True, 'property_type': 'House', 'room_type': 'Entire home/apt',
'accommodates': 9.0, 'bathrooms': 2.5, 'bedrooms': 4.0, 'beds': 4.0, 'bed_type': 'Real Bed',
```

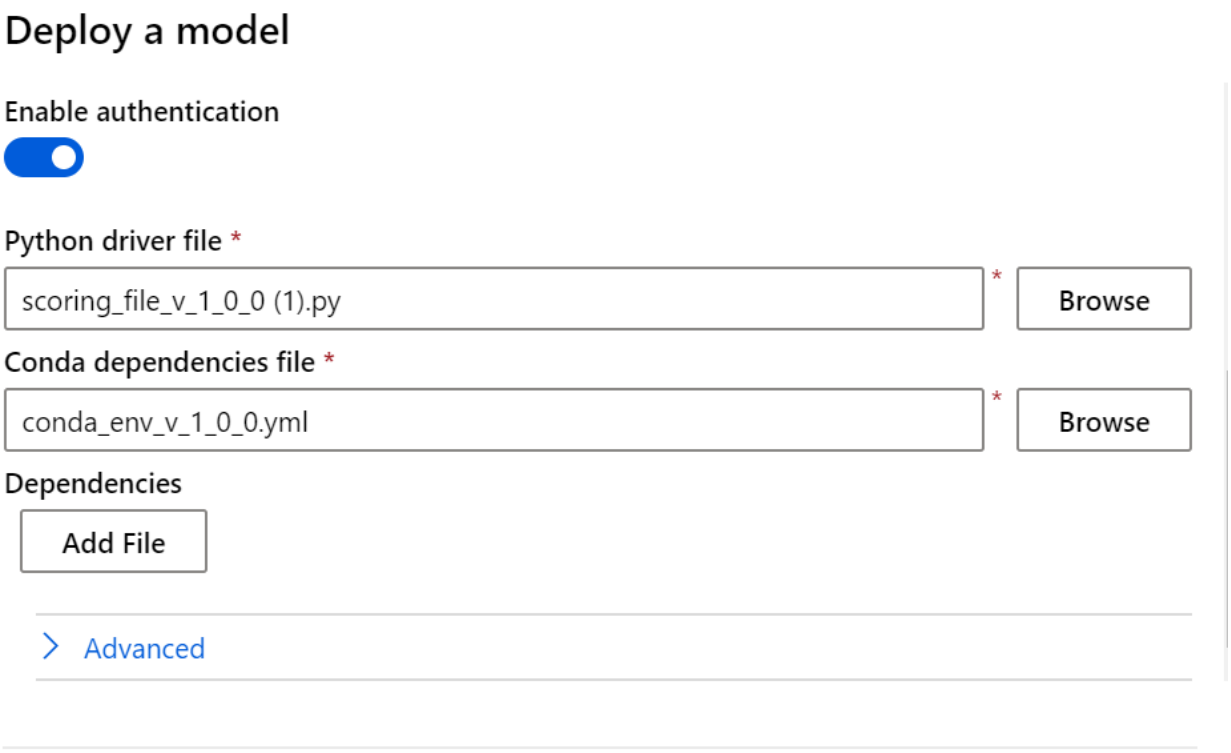
```
'square_feet': None, 'price': 296.0, 'weekly_price': None, 'monthly_price': None,
'security_deposit': 500.0, 'cleaning_fee': 250.0, 'guests_included': 8.0, 'extra_people': 25.0,
'minimum_nights': 30.0, 'maximum_nights': 1000.0, 'minimum_minimum_nights': 30.0,
'maximum_minimum_nights': 30.0, 'minimum_maximum_nights': 1000.0,
'maximum_maximum_nights': 1000.0, 'minimum_nights_avg_ntm': 30.0,
'maximum_nights_avg_ntm': 1000.0, 'calendar_updated': '5 days ago', 'has_availability': True,
'availability_30': 25.0, 'availability_60': 55.0, 'availability_90': 84.0, 'availability_365': 84.0,
'calendar_last_scraped': '2019-09-22T00:00:00.000Z', 'number_of_reviews': 28.0,
'number_of_reviews_ltm': 8.0, 'first_review': '2008-09-15T00:00:00.000Z', 'last_review': '2019-
08-30T00:00:00.000Z', 'review_scores_rating': 100.0, 'review_scores_accuracy': 10.0,
'review_scores_cleanliness': 10.0, 'review_scores_checkin': 10.0,
'review_scores_communication': 10.0, 'review_scores_location': 10.0, 'review_scores_value':
10.0, 'requires_license': True, 'instant_bookable': False, 'is_business_travel_ready': False,
'require_guest_profile_picture': False, 'require_guest_phone_verification': False,
'calculated_host_listings_count': 2.0, 'calculated_host_listings_count_entire_homes': 2.0,
'calculated_host_listings_count_private_rooms': 0.0,
'calculated_host_listings_count_shared_rooms': 0.0, 'reviews_per_month': 0.21}])
output_sample = np.array([0])


def init():
    global model
    # This name is model.id of model that we want to deploy deserialize the model file back
    # into a sklearn model
    model_path = Model.get_model_path(model_name = 'AutoML5d0435c1034')
    model = joblib.load(model_path)


@input_schema('data', PandasParameterType(input_sample))
@output_schema(NumpyParameterType(output_sample))
def run(data):
    try:
```

```
        result = model.predict(data)
        return json.dumps({"result": result.tolist()})
    except Exception as e:
        result = str(e)
        return json.dumps({"error": result})
```

Go to model and deploy the best model



Upload conda and scoring file



Weka Attribute Selection

Attribute evaluator	Classifier	Search Method	# of attribute selected	Merit of the best subset
WrapperSubsetEvaluate	J48	Best fit - Backward	4	0.774
WrapperSubsetEvaluate	J48	Best fit - Forward	4	0.774
WrapperSubsetEvaluate	Naïve Bayes	Best fit - Backward	4	0.741
WrapperSubsetEvaluate	Naïve Bayes	Best fit - Bi directional	4	0.741
WrapperSubsetEvaluate	Ibk	Best fit - Backward	4	0.782
WrapperSubsetEvaluate	Ibk	Best fit - Forward	5	0.78
WrapperSubsetEvaluate	SMO	Best fit - Backward	4	0.75

WrapperSubsetEvaluate	SMO	Best fit - Forward	5	0.78

Attribute evaluator	Classifier	Search Method	# of attribute selected	Merit of the best subset
ClassifierSubsetEval	J48	Best fit - Backward	6	0.835
ClassifierSubsetEval	J48	Best fit - Forward	5	0.842
ClassifierSubsetEval	Naïve Bayes	Best fit - Backward	5	0.881
ClassifierSubsetEval	Naïve Bayes	Best fit - Forward	5	0.885
ClassifierSubsetEval	Ibk	Best fit - Backward	2	1
ClassifierSubsetEval	Ibk	Best fit - Forward	1	1
ClassifierSubsetEval	SMO	Best fit - Backward	3	0.996

Accuracy classifier with the attribute selected from the evaluator:

Variation of selected Attribute	SVM (SMO)	Decision Tree (J48)	kNN (Ibk)	Naïve Bayes	Average Accuracy
Iteration 1	77.2931	65.1379	72.5465	65.1379	70.02885
Iteration 2	72.1000	68.2809	78.1719	73.6049	73.03943
Iteration 3	65.1000	68.2809	77.9346	72.0013	70.82920
Iteration 4	69.1200	76.0103	72.1937	53.6241	67.73703
Iteration 5	85.1000	76.0207	72.1937	53.6883	71.75068
Iteration 6	76.3000	68.2809	69.5959	70.7505	71.23183
Iteration 7	75.9000	68.2809	70.0128	70.4298	71.15588
Iteration 8	72.1000	68.2809	75.4971	67.1388	70.75420
Iteration 9	68.2809	72.9314	69.628	52.9827	65.95575
Iteration 10	64.1000	69.5959	68.2809	67.5754	68.48407