Walmart Sales Forecasting

A CRISP-DM Model

Data Governance

Table of Contents

Executive summary

As part of Data Science workflow understanding of business and building prediction models one of the most critical functions which plays an important role towards closing this loop effectively. I planning to explore this with dataset from Walmart, and will show some insight, analyze anomalies and patterns, will build model for sales forecasting and explain all the steps of it with help of CRISP-DM methodology.

1. Business Understanding

1.1 Determine Business Objectives

Background

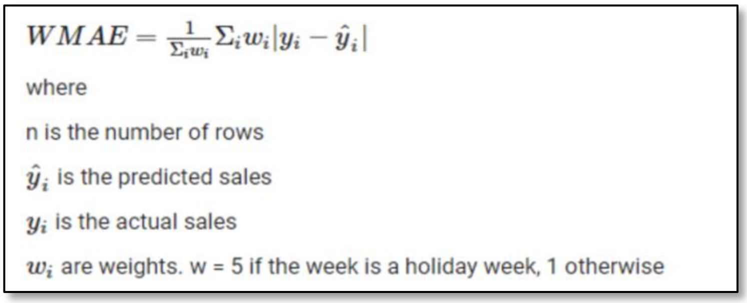
Walmart is a renowned retail corporation that operates a chain of hypermarkets. Here, Walmart has provided a data combining of 45 stores including store information and monthly sales. The data is provided on weekly basis. Walmart tries to find the impact of holidays on the sales of store. For which it has included four holidays' weeks into the dataset which are Christmas, Thanksgiving, Super bowl, Labor Day. Here we are owing to Analyze the dataset given. before doing that, let me point out the objective of this analysis. (9)

Business Objectives

Our Main Objective is to predict sales of store in a week. As in dataset size and time related data are given as feature, so analyze if sales are impacted by time-based factors and space-based factor. Most importantly how inclusion of holidays in a week soars the sales in store? (9)

Business Success Criteria

A Machine Learning model like, a regression model can provide robust prediction given the dataset satisfies its linearity assumptions. Furthermore, machine learning forecasting is not a black box; the influence of model inputs can be weighed and understood so that the forecast is intuitive and transparent. Machine Learning models can also be updated and become adaptable to the changes in dataset. And also, through machine learning help, relation between markdown events and weekly sales can be utilize in correct manner using machine learning model. A machine learning model requires a metric that will check weather model is performing good or bad. This metric is called Weighted Mean Absolute Error. (9)



$$WMAE = \frac{1}{\Sigma_i w_i}\Sigma_i w_i|y_i - \hat{y}_i|$$

where

n is the number of rows

$\hat{y}_i$ is the predicted sales

y_i is the actual sales

w_i are weights. w = 5 if the week is a holiday week, 1 otherwise

It is obvious that our aim is to lower this score to as much extent as possible along with predicting weekly sales. We will use this performance metric in regression models. (9)

With the accurate prediction company can:

- Determine seasonal demands and take action for this
- Protect from money loss because achieving sales targets can have a positive effect on stock prices and investors' perceptions
- Forecast revenue easily and accurately
- Manage inventories
- Do more effective campaigns

1.2 Assess Situation

Inventory of Resources

The following list of tangible and intangible resources is provided to ensure a smooth and proper undertaking of this project:

Intangible Assets (Resources) – Personnel:

- *Data Analysts* - (a team of two):
- *Data Scientist* – to oversee the successful completion of this project
- *IT staff member* – to oversee software, hardware, and internet connectivity issues as they arise.
- *Legal and compliance team* – to review business project contract
- *Accounting* – for expense tracking and invoices
- *Project Manager* (10)

Tangible Assets (Resources)

- *Data: Kaggle datasets*
- *Software: Microsoft Excel* - to handle the raw examination of the .csv file (as backup).
- *Software: Anaconda Navigator* - installed and equipped with Jupyter Notebooks to load in the .csv file for analysis.
- *Software: Windows 10* - running on a PC with at least 16 gigabytes of ram.
- *Zendesk:* - to log IT related support issues should they arise. (10)

Requirements, Assumptions & Constraints

Purpose, and effect of this project is to ensure comprehensibility and digestibility across many vast user profiles with a basic business vocabulary. To this end, will inclusively and holistically involve staff members from each department mentioned in the inventory of resources. However, to narrow the focus of the final product, understanding trends and patterns should not be a complex undertaking that only people well-versed in the language of data analytics need understand. For this reason, the resulting models will be relegated to high-level graphs and charts that show overall trends. Furthermore, the ensuing models will be evaluated iteratively, until deployment is finalized, at which point the VP of Marketing and the Chief Financial Officer will work on ensuring repeatability and implementation. Data-warehousing should not be disregarded. All end users will be provided with two-factor authentication to ensure an extra layer of protection. More importantly, we must not ignore the legal constraints. There are the obvious technological limitations of insufficient processing speed hampering performance on inadvertently overfitted dataset. A slower processing speed, though sometimes unforeseen, should not compromise data integrity. (10)

Risk & Contingencies

Data governance helps to ensure that data is usable, accessible and protected. Effective data governance leads to better data analytics, which in turn leads to better decision making and improved operations support. Further, it helps to avoid data inconsistencies or errors in data, which lead to integrity issues, poor decision making, and a variety of organizational problems. (11) With any business endeavor there exists a certain amount of risk as it pertains to the resources of money, human capital, and labor hours:

- *Unexpected system outages* (i.e., internet connectivity).

In the rare and unlikely event of an internet outage, especially when working from a home office, it is crucial to coordinate with IT the procurement of mobile wireless hotspots. IT will oversee and monitor any issues pertaining to gateway security, and privacy, but this must be assured due to the sensitive nature of this project.
Zendesk support tickets will be used to log IT related issues.

- *Security risks* (i.e., privacy, data integrity).

In the event of a data breach or leak of information, all parties must be made aware of the policies and procedures.

- *Staffing resources unavailable* (i.e., temporary or permanent reduction in staff hours dedicated to this project).

With any project, a dedicated staff may experience untimely and unforeseen emergencies leading to a reduction in hours. For this reason, we have a team of two data analysts. Both will be cross trained in each other's daily routines and handling of tasks in the event of absence by one or the other. Especially now with the pandemic, untimely absenteeism needs to be taken with a grain of salt. (10)

- *Limited productivity due to distributed teams*. Staff deliverables between analysts and data scientists may not be met or produced in a timely manner.

To maximize workflow efficiency, proper communication channels will be established from the onset. For example, the team will utilize Microsoft Teams in lieu of face-to-face meetings, where such communications and collaboration will be facilitated by the project manager. (10)

Terminology

Anaconda Navigator: "a desktop GUI that comes with Anaconda Individual Edition. It makes it easy to launch applications and manage packages and environments without using command-line commands." (12)
Jupyter Notebook: "an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more." (13)
Linear Regression: "regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables)." (14)
Data Analytics: "the science of analyzing raw data in order to make conclusions about that information. Many of techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption." (15)

Machine Learning: " an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed." (16)
Data mining: " process of discovering interesting patterns from massive amounts of data." (17)
CPI: "measure that examines the weighted average of prices of a basket of consumer goods and services." (18)
Weighted Mean Absolute Error: " measure of prediction accuracy of forecasting method. (19)

Cost & Benefits

Comprehensive analysis of the true cost of data management and analytic operations will consist:

- The price of the software tools that will use to manage and analyze data.
- The cost of storage infrastructure data.
- The cost in terms of staff time that company's data engineers spend managing data.
- Inefficient data integration

Deriving value from data typically requires to transform and integrate it. Data integration solution involves a lot of manual effort, redundancy or other types of inefficiency, it could be a major source of costs in the form of wasted staff time and unnecessary infrastructure. (20)

- Networking data costs.

Whether we are only transferring data locally on internal network, or we are moving data between a cloud and on-premise infrastructure, the networking devices and bandwidth that we will need have a price.

- Low-quality data.

While avoiding data quality problems entirely may not be possible, correcting data quality issues via automated tools can help to minimize the costs incur due to low data quality.

- Unnecessary backups

We should always back up data. However, if we are backing up data more frequently than we require or storing more copies than need, we are not operating in the most cost-effective way possible.

- Unqualified employees

The lack of knowledge on the part of employees could bloat costs. Avoid this cost by turning your employees into data scientists. (20)

1.3 Determine Data Mining Goals

Data Mining Goals

To enact an effective sale forecast strategy, a strong inverse linear correlation between store unit and sales volume must be established. This can be accomplished within the scope and context of a supervised regression model. The initial exploratory data analysis process will determine and unveil the independent (predictor) variables, distinguishing them from the dependent (response) variable of what we are trying to predict (sales). (21)

Data Mining Success Criteria

Statistical significance from this sample size will be established under a confidence level which has shown less mean absolute error as possible, at least less 900. If the result yields in favor of statistical significance, we can conclude that the results have occurred due to chance, and not any kind of pre-processing or manipulation of the model. Rejecting or failing to reject our initial claim hinges on this framework. Testing the model in and of itself is a goal, attribution, and testament to a successful analytics endeavor. Realizing that our challenge may extend beyond a traditional linear regression model, we are prepared to explore alternative logistical models, which we will discuss in greater detail. (21)

1.4 Produce Project Plan

Project Plan

TASK	DURATION IN DAYS
PHASE 1 Digest the metrics of Walmart's current sales strategy and its impact on annual revenue. Draft and discuss risks, contingencies, and terminology. Refine the economic model to account for first and second order conditions. Research last 2 years' revenue metrics year over year, and month over month to identify sales patterms. Coordinate meeting with IT, the two data analysts, and data scientist to ensure data software is made accessible without any restrictions.	
Business Understanding	10
Data Understanding	7
Data Preparation	2
Data Understanding	3
PHASE 2 Exploratory Data Analysis (EDA). Set up a basic model in python. Evaluate model based on strong factors of strong predictive metrics. Model will be trained and tested for accuracy and precision. Deploy optimal pricing model.	
Data Preparation	2
Data Understanding	5
Modeling	15
Evaluation	TBD
Deployment	TBD

Initial Assessment of Tools & Techniques

For this project, we used techniques involve with data mining, such as:

- Data cleaning: To remove noise and inconsistent data.
- Data integration: where multiple data sources may be combined.
- Data selection: Where data relevant to the analysis task are retrieved from the database
- Data transformation: Where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations.
- Data mining: An essential process where intelligent methods are applied to extract data patterns.
- Pattern evaluation: To identify the truly interesting patterns representing knowledge based on interestingness measures.
- Knowledge presentation: Where visualization and knowledge representation techniques are used to present mined knowledge to users. (22)

2. Data Understanding

2.1 Collect Initial Data

Initial Data Collection Report

“Data collection” is the initial step before starting to analyze the patterns or useful information in data. The data which is to be analyzed must be collected from different valid sources. In this project we will use “quantitative data” which is in numerical forms and can be calculated using different scientific tools and sampling data. We planning to use secondary data which has already been collected and will reused again for valid purpose. This type of data is previously recorded from primary data and it has two types of sources named internal source and external source. (23)
Internal source: These types of data can easily be found within the organization such as market record, a sales record, transactions, customer data, accounting resources, etc. The cost and time consumption are less in obtaining internal sources.
External source: The data which can’t be found at internal organizations and can be gained through external third-party resources is external source data. The cost and time consumption are more because this contains a huge amount of data. Examples of external sources are Government publications of Weather, planning commission, Fuel cost by date, Government publications of Holidays. (23)

2.2 Describe Data

Data Description Report

In this project implementation of data policies in an operational area and monitors the data quality as well will mainly task for our data steward, who are responsibility for a subset of data. However, our senior managers in the organizations are responsible for specifying the organization’s requirements on data and on data quality. Responsible for data creation and maintenance and also accountable for the data definition in specific areas of responsibility which differs according to the role and data requirement.(23)

2.3 Explore Data

Data Exploration Report

We will carry out the process called Exploratory Data Analysis, an approach to analyzing data sets to summarize their main characteristics.

- There are 45 stores and 81 department in data. Departments are not same in all stores.
- Although department 72 has higher weekly sales values, on average department 92 is the best. It shows us, some departments has higher values as seasonal like Thanksgiving. It is consistant when we look at the top 5 sales in data, all of them belongs to 72th department at Thanksgiving holiday time.
- Although stores 10 and 35 have higher weekly sales values sometimes, in general average store 20 and store 4 are on the first and second rank. It means that some areas has higher seasonal sales.
- Stores has 3 types as A, B and C according to their sizes. Almost half of the stores are bigger than 150000 and categorized as A. According to type, sales of the stores are changing.

- As expected, holiday average sales are higher than normal dates.
- Christmas holiday introduces as the last days of the year. But people generally shop at 51th week. So, when we look at the total sales of holidays, Thankgiving has higher sales between them which was assigned by Walmart.
- Year 2010 has higher sales than 2011 and 2012. But, November and December sales are not in the data for 2012. Even without highest sale months, 2012 is not significantly less than 2010, so after adding last two months, it can be first.
- It is obviously seen that week 51 and 47 have higher values and 50-48 weeks follow them. Interestingly, 5th top sales belongs to 22th week of the year. This results show that Christmas, Thankgiving and Black Friday are very important than other weeks for sales and 5th important time is 22th week of the year and it is end of the May, when schools are closed. Most probably, people are preparing for holiday at the end of the May.
- January sales are significantly less than other months. This is the result of November and December high sales. After two high sales month, people prefer to pay less on January.
- CPI, temperature, unemployment rate and fuel price have no pattern on weekly sales.(24)

2.4 Verify Data Quality

Data Quality Report

From the pivot table, it is obviously seen that there are some wrong values such as there are 0 and minus values for weekly sales. But sales amount can not be minus. Also, it is impossible for one department not to sell anything whole week. Rarely, some of store may was close of this or those reason like force major. Sometime no records for the sales simple can be as result of human mistakes when data was not collected in correct moaner or was not store in correct source. So, we will change these values. 1358 rows in 421570 rows means 0.3%, so we can ignore these rows which contains wrong sales values. (25)

3. Data Preparation

3.1 Select Data

Data Set

The dataset comes from an American retail organization, Walmart Inc. It comprises information from 45 Walmart division stores, from 2010 to 2012 primarily centered around their deals on a week after week premise. Each section has properties as takes after: the related store (recorded as a number), the comparing division (81 offices, each entered as a number), the date of the beginning day in that week, departmental week after week deals, the store measure, and a Boolean esteem indicating on the off chance that there's a major occasion within the week. The major occasions being one of Thanksgiving, Labor Day, Christmas or Easter. Together with the previously mentioned qualities may be a parallel set of highlights for each section counting Customer Cost List, unemployment rate, temperature, fuel cost, and special markdowns. (26)

Data Set Description

There are four datasets provided by Walmart to build a predictive model.
Stores.csv - this file contains anonymized information about the 45 stores, indicating the type and size of the store:
Store: stores numbered from 1 to 45
Type: store type has been provided, there are 3 types — A, B and C.
Size: stores size has provided
Train.csv - this is the historical training data, which covers 2010–02–05 to 2012–11–01. Within this file you will find the following fields:
Store: the store number
Dept: the department number
Date: the week
Weekly_Sales: sales for the given department in the given store,
IsHoliday: whether the week is a special holiday week
Test.csv - This file is identical to train.csv, except we have withheld the weekly sales. We x§
Dept: the department number
Date: the week
IsHoliday: whether the week is a special holiday week
Features.csv - this file contains additional data related to the store, department, and regional activity for the given dates. It contains the following fields:
Store: the store number
Date: the week
Temperature: average temperature in the region
Fuel_Price: cost of fuel in the region
MarkDown1–5: anonymized data related to promotional markdowns that Walmart is running. MarkDown data is only available after Nov 2011 and is not available for all stores all the time. Any missing value is marked with an NA.
CPI: the consumer price index
Unemployment: the unemployment rate
IsHoliday: whether the week is a special holiday week

Rationale For Inclusion/Exclusion

All our data in dataset can be used for future business analyze, it provided from open source for future work. Data is sufficient for exploring insight in business and to build predictive sales forecast model.

3.2 Clean Data

Data Cleaning Report

- The data has no too much missing values. All columns was checked.
- We choose rows which has higher than 0 weekly sales. Minus values are 0.3% of data.
- Null values in markdowns changed to zero. Because, they were written as null if there were no markdown on this department. (25)

3.3 Construct Data

Derived Attributes

We need to see differences between holiday types. So, I create new columns for 4 types of holidays and fill them with boolean values. If date belongs to this type of holiday it is True, if not False. As well, we will change Date to Datetime (month, year) and creating new columns for better business analyzes exploration. (25)

Generated Records

Imbalanced data refers to those types of datasets where the target class has an uneven distribution of observations, i.e one class label has a very high number of observations and the other has a very low number of observations. In our case we can’t see such problem, however in such cases we can use Synthetic Minority Oversampling Technique or SMOTE is another technique to oversample the minority class. Simply adding duplicate records of minority class often don’t add any new information to the model. In SMOTE new instances are synthesized from the existing data.

3.4 Integrate Data

Merged Data

"Merging” datasets are the process of bringing all datasets together into one, and aligning the rows from each based on common attributes or columns. In this case we will merge all three different sets ("Store", "Date"- from features and "Store" - from Store), will remove duplicated columns such as "IsHoliday_y". After merging data, we will store it as new csv file.(27)

3.5 Format Data

Formatted Data

We rely on AI/ML models to gain insights and make predictions about the future. Data quality is directly related to the effectiveness of AI/ML models, as high-quality data means that knowledge about the past is less biased. Consequently, this leads to better forecasts.
As the image above suggests, low-quality data or data scarcity leads workers to spend more effort on tasks that do not add value. This is because without AI/ML models, every task must be done manually regardless of its yield. So, ensuring data quality is of great importance in guaranteeing the efficiency of business operations. With this idea lets will change IsHoliday column with ‘False’ to be 0 and ‘True’ to be 1. Similarly, convert ‘Types’ of the store to numeric values. Missing Value like in ‘Markdown’, Imputing it with Zero(No Markdown). We can safely fill all missing values with zero. so it definety will improve our data quality and as resolt score.(28)

4. Modeling

4.1 Select Modeling Technique

Modeling Technique

For this project we are planning to use some baseline model and time series models such as KNN Regressor, Decision Tree Regressor, RandomForestRegressor, Auto-ARIMA, Exponential Smoothing. We will check all of the to identify which one will produce the best result. Secondly our data is that it is not stationary. To make data more stationary taking difference,log and shift techniques will applied.

Modeling Assumptions

- Assumption of no formal distributions. Being a non-parametric model, it can handle skewed and multi-modal data.
- It may have an assumption that encoded integer value for each variable has ordinal relation.
- Initially, whole training data is considered as root.
- Records are distributed recursively on the basis of the attribute value.
- The data is in feature space, which means data in feature space can be measured by distance metrics such as Manhattan, Euclidean, etc.
- Each of the training data points consists of a set of vectors and a class label associated with each vector. (29)

4.2 Generate Test Design

Test Design

The model is estimated in metric weighted mean absolute error (WMAE), so we will to use function to calculate it this error. The lowest error will be our best performance in this stage. When we will observe the lowest error that the signs are stable.

4.3 Build Model

Parameter Settings

Firstly, among all the models we need to choose the model with the best performance. Secondly to improve our performance we will implement some experiments and decisions to find out the best results as:

- model without divided holiday columns
- model without month column
- model whole data
- model whole data with feature selection
- model whole data with feature selection without month
- model without divided holiday columns (30)

Models

Model versioning in a way involves tracking the changes made to an ML model that has been previously built. Put differently, it is the process of making changes to the configurations of an ML Model. From another perspective, we can see model versioning as a feature that helps Machine Learning Engineers, Data Scientists, and related personnel create and keep multiple versions of the same model. There are many outstanding tools in these days I will suggest data version control tool such as Neptune. (31)

Neptune AI is an MLOps tool that allows you to track and compare model versions. It is primarily built for ML and data science teams to run and track experiments. The platform allows you to store, version, query, and organize models and associated metadata. Specifically, you can store and version things like your training data, code, environment configuration versions, parameters, hyperparameters and evaluation metrics, testset predictions, etc. (31)

Model Description

Once a machine learning model is trained and deployed in production, there are two approaches to monitor its performance degradation: ground truth evaluation and input drift detection:

Performance degradation.
In simpler terms, our model is no longer modelling the outcome that it used to model. This causes problems because the predictions become less accurate as time passes. One solution is referred to as manual learning. Here, we provide the newly gathered data to our model and re-train and re-deploy it just like the first time we build the model. If you think this sounds time-consuming, you are right. Moreover, the tricky part is not refreshing and retraining a model, but rather thinking of new features that might deal with the concept drift. And in my opinion is the best solution is to build productionized system in such a way that you continuously evaluate and retrain your models. The benefit of such a continuous learning system is that it can be automated to a large extent, thus reducing (the human labor) maintained costs. (32)

Drift detection.
Data-drift is defined as a variation in the production data from the data that was used to test and validate the model before deploying it in production. There are many factors that can cause data to drift one key factor is the time dimension. When data drifts it is not identified on time, the predictions will go wrong, our business decision will be taken based on the predictions may have a negative impact. It is important to build a repeatable process to identify data drift, define thresholds on drift percentage, configure pro-active alerting so that appropriate action is taken. (33)

4.4 Assess Model

Model Assessment

According to a test strategy which we build we can compare evaluation results and create ranking of our results with respect to success and evaluation criteria.

1. Exponential Smoothing model whole data with feature selection, WMAE 821
2. RandomForestRegressor model whole data with feature selection, WMAE 1801
3. RandomForestRegressor model whole data with feature selection without month, WMAE 2093
4. RandomForestRegressor model whole data, WMAE 2450
5. RandomForestRegressor model without month column, WMAE 5494
6. RandomForestRegressor model without divided holiday columns, WMAE 5850(30)

Revised Parameter Settings

An artifact can be an archive files that includes the following structural elements:

- Compilation output for one or more of our modules
- Libraries included in module dependencies
- Collections of resources (web pages, images, descriptor files, and so on)
- Other artifacts
- Individual files, directories and archives(34)

5. Evaluation

5.1 Evaluate Results

Assessment of Data Mining Results

We can conclude that the results have occurred due to kind of pre-processing or manipulation of the model, and not chance. It has been shown that, all our data mining goals are achieved and perfume great results. And also, through machine learning model help, relation between markdown events and weekly sales can be utilize in correct manner. With the prediction company can determine seasonal demands, protect from money loss, forecast revenue, manage inventories and do more effective campaigns.

Approved Models

The best performance we can clearly identify in Exponential Smoothing model which contain whole data with feature selection, WMAE 821. According to sales amounts, if we can take our average sales and take percentage of WMAE 821 errors, it gives 4-5% roughly. Which is great performance for a business objective of our aim. (30)

5.2 Review Process

Review of Process

Metric which we use to check weather model is performing good or bad is Weighted Mean Absolute Error. WMAE matched the problem at hand, which means understanding the limits and trade-offs of the metric (the mathematics side) and their impact on the optimization of the model (the business side). The most valid metric among suggested one such as Mean Squared Error Metric and Mean Squared Error Metric. (35)

5.3 Determine Next Steps

List of Possible Actions

For a future work in this predictive modeling problem we can use Transfer learning. TL is a ML technique where a model trained on one task is re-purposed on a second related task, optimization that allows rapid progress or improved performance when modeling the second task. We can transfer learning in this predictive modeling problems with two common approaches as Develop Model, Pre-trained Model. Transfer learning is an optimization, a shortcut to saving time or getting better performance which will defiantly have great input our business goals as well. There are few benefits to use Transfer learning for future work:

- *Higher start.* The initial skill (before refining the model) on the source model is higher than it otherwise would be.
- *Higher slope.* The rate of improvement of skill during training of the source model is steeper than it otherwise would be.
- *Higher asymptote.* The converged skill of the trained model is better than it otherwise would be. (36)

Decision

Explainable AI has strategic value for business leaders it can accelerate AI adoption, enable accountability, provide strategic insights, and ensure ethics and compliance. As explain ability helps build the trust and confidence of stakeholders in the ML, it increases the adoption of AI systems in the organization providing it a competitive advantage. Explain ability will give confidence to our organizational leaders to accept the accountability for the AI systems in their business as it provides them a better understanding of the systems' behavior and potential risks. This promotes greater executive buy-in and sponsorship for AI projects. With the support of key stakeholders and executives for AI, the organization will be better positioned to foster innovation, transformation, and developing next-generation capabilities.

Explainable models can also help provide valuable insights into key business metrics such as sales forecasting models, used to predict sales and plan inventory. If the forecasting models can also show how the key factors like price, promotion, competition, etc., contribute to sales forecast, that information can be used to boost sales. (37)

Responsible AI is governance framework that documents how our organization can deploy AI in manner that is ethical and legal. It marries the need for AI to be used for good and the opportunity it provides businesses, governments and broader society. Part of this is the role of human oversight built on ethics and data governance or freedoms. Responsible AI is not necessarily about humans overseeing every single decision that our algorithm makes, but ensuring that they are equipped to provide unbiased, ethical and restrictive outcomes. (38)

6. Deployment

6.1 Plan Deployment

Deployment Plan

Version Management is a means to effectively track and control changes to a collection of related entities. It is a very important tool within an overall life cycle management strategy for information solutions. Traditionally Version Management have relied upon a central repository for storing and tracking changes to managed entities. For these systems individual users check out portions of the central repository into their own local workspaces where they make the desired changes. Once the changes are ready, they commit those changes back into the central repository. (39)

Key benefits of Version Management:

- Organized, coordinated management of changes to assets by one or many individuals, some of whom may be geographically dispersed.
- Organized, coordinated management of changes to assets for emergency hot-fixes, routine maintenance, upgrades, and new features with potentially overlapping development timeframes.
- A reliable master copy of what assets are currently in production.
- A reliable master copy of assets from which to build and/or configure the production environment.
- Reliable copies of previous production versions of assets. (39)

6.2 Plan Monitoring & Maintenance

Monitoring & Maintenance Plan

After deploying our model monitoring and maintenance will take place. One of the most important aspects in this stage is accuracy. In this stage once accuracy not performing as we desire maintenance need to take a place in role as update the model or set up new data mining project.

6.3 Produce Final Report

Final Report

At the end we need to be aware and prepared that model can act in not proper way in one day. All business decision which was taken need to be reconsider, model checked from scratches. Our main aim will to identify bug or human mistakes which transfer our model to such this consequence. Reobserve problem and placement it in main stage to prevent such cases.

Final Presentation

The data which we analyze was collected from different valid sources and under following terms of use:

- *Notice:* users should be given notice when their data is being collected.
- *Purpose:* data should only be used for what you say you will use it for.
- *Consent:* user data should not be shared without your users' consent.
- *Security:* collected data should be kept secure.
- *Disclosure:* users should be informed about who is collecting their data. (40)

Data accuracy has many benefits for any business and could save you from making an incredibly poor business decision. For the highest data accuracy, businesses should set data accuracy standards that must be met by any data that is processed into the system. Data accuracy standards include data profiling, data monitoring, data linking and matching. Having these standards ensures that all data used conforms to predefined standards which will improve data quality. (41)

6.4 Review Project

Experience Documentation

We need have collaboration with organization which can help and focus on certification and compliance, provide a system component that's part of the overall data governance framework. This option can support to our process and provide for us more time and futures which could be utilize for other matter.

Conclusion

Most of the shopping malls / shopping centers plan to attract the customers to the store and make profit to the maximum extent by them. Once the customers enter the stores they are attracted then definitely they shop more by the special offers and obtain the desired items which are available in the favorable cost and satisfy them. If the products as per the needs of the customers, then it can make maximum profit the retailers can also make the changes in the operations, objectives of the store that cause loss and efficient methods can be applied to gain more profit by observing the history of data the existing stores a clear idea of sales can be known like seasonality trend and randomness. The advantage of forecasting is to know the number of employees should be appointed to meet the production level. Sales drop is bad thing forecasting sales helps to analyze it and it can overcome through the sales drop to remain in the competition forecast plays a vital role.

References

(1) Artificial intelligence in business [Online]. Available: https://www.nibusinessinfo.co.uk/content/business-benefits-artificial-intelligence [Accessed: March 11, 2022].

(2) AI Project Life Cycle: Important Stages and Details [Online]. Available: https://www.maxinai.com/resources/understanding-ai-project-cycle-important-stages-details [Accessed: March 11, 2022].

(3) (July 16, 2020), The Role of Quality Assurance in Artificial Intelligence [Online]. Available: https://appen.com/blog/quality-assurance-in-ai/ [Accessed: March 11, 2022].

(4) Data Lakes vs. Data Warehouses [Online]. Available: https://www.datacamp.com/blog/data-lakes-vs-data-warehouses [Accessed: March 11, 2022].

(5) The three-pillar approach to cyber security: Data and information protection [Online]. Available: https://www.dnv.com/article/the-three-pillar-approach-to-cyber-security-data-and-information-protection-165683 [Accessed: March 11, 2022].

(6) Shrutika M. & A. R. Tripathi (02 July 2021), AI business model: an integrative business approach [Online]. Available: https://innovation-entrepreneurship.springeropen.com/articles/10.1186/s13731-021-00157-5 [Accessed: March 11, 2022].

(7) Claire (12 March 2021), How Big Data and AI Are Driving Business Innovation in 2021 [Online]. Available: https://www.virtualedge.org/how-big-data-and-ai-are-driving-business-innovation/ [Accessed: March 11, 2022].

(8) Amazon Web Services (5 October 2021), Unlocking The Power of Data: Making Better Business Decisions [Online]. Available: https://www.forbes.com/sites/amazonwebservices/2021/10/05/unlocking-the-power-of-data-making-better-business-decisions/?sh=72be01d959dd [Accessed: March 11, 2022].

(9) Aayush G (26 May 2021), Walmart Recruiting - Store Sales Forecasting [Online]. Available: https://medium.com/geekculture/walmart-recruiting-store-sales-forecasting-b8b2f4cf19b1 [Accessed: March 11, 2022].

(10) Shpaner L. (24 February 2021), Walmart Sales Forecasting [Online]. Available: https://www.leonshpaner.com/projects/post/walmart_price_model/ [Accessed: March 11, 2022].

(11) Watts S. (16 April 2020), What Is Data Governance? Why Do I Need It? [Online]. Available: https://www.bmc.com/blogs/data-governance/ [Accessed: March 11, 2022].

(12) Anaconda Navigator [Online]. Available: https://docs.anaconda.com/anaconda/navigator/index.html [Accessed: March 11, 2022].

(13) JupyterLab: A Next-Generation Notebook Interface [Online]. Available: https://jupyter.org [Accessed: March 11, 2022].

(14) Brownlee J. (25 March 2016), Linear Regression for Machine Learning [Online]. Available: https://machinelearningmastery.com/linear-regression-for-machine-learning/ [Accessed: March 11, 2022].

(15) Frankenfield J. (04 September 2021), Data Analytics [Online]. Available: https://www.investopedia.com/terms/d/data-analytics.asp [Accessed: March 11, 2022].

(16) Expert.ai Team (06 May 2020), What Is Machine Learning? A Definition. [Online]. Available: https://www.expert.ai/blog/machine-learning-definition/ [Accessed: March 11, 2022].

(17) Data Mining [Online]. Available: https://www.sas.com/en_ae/insights/analytics/data-mining.html [Accessed: March 11, 2022].

(18) Fernando J. (10 March 2022), Consumer Price Index [Online]. Available: https://www.investopedia.com/terms/c/consumerpriceindex.asp [Accessed: March 11, 2022].
(19) (06 February 2021), WMAPE [Online]. Available: https://en.wikipedia.org/wiki/WMAPE [Accessed: March 11, 2022].
(20) Tozzi C. (8 June 2020), The Hidden Costs of Big Data [Online]. Available: https://www.precisely.com/blog/big-data/the-hidden-costs-of-big-data [Accessed: March 11, 2022].
(21) Thakur S. (10 June 2016), Different Goals of Data Mining [Online]. Available: https://whatisdbms.com/different-goals-of-data-mining/ [Accessed: March 11, 2022].
(22) Santaella C.J.G. Data Mining Techniques and Machine Learning Model for Walmart Weekly Sales Forecast [Online]. Available: https://prcrepository.org/xmlui/bitstream/handle/20.500.12475/174/FA-19_Articulo%20Final_Jose%20Santaella.pdf?sequence=1&isAllowed=y [Accessed: March 11, 2022].
(23) (10 September 2020), Different Sources of Data for Data Analysis [Online]. Available: https://www.geeksforgeeks.org/different-sources-of-data-for-data-analysis/?ref=rp [Accessed: March 11, 2022].
(24) Gumusbas E. (06 May 2020), Project4_Store_Sales_Forecasting [Online]. Available: https://github.com/ezgigm/Project4_Store_Sales_Forecasting/blob/master/STEP1_Cleaning_and_EDA.ipynb [Accessed: March 11, 2022].
(25) Gumusbas E. (31 August 2020), Walmart Store Sales Forecasting [Online]. Available: https://github.com/ezgigm/Project4_Store_Sales_Forecasting [Accessed: March 11, 2022].
(26) Aldukhayni A. (22 December 2020), Walmart-recruiting-store-sales-forecasting [Online]. Available: https://medium.com/analytics-vidhya/walmart-recruiting-store-sales-forecasting-ef275e8d5d4e [Accessed: March 11, 2022].
(27) Lee A. (02 March 2019), Why and How to Use Merge with Pandas in Python [Online]. Available: https://towardsdatascience.com/why-and-how-to-use-merge-with-pandas-in-python-548600f7e738 [Accessed: March 11, 2022].
(28) Görkem G. (10 January 2022), Data Quality Assurance: What it is & Best Practices [Online]. Available: https://research.aimultiple.com/data-quality-assurance/ [Accessed: March 11, 2022].
(29) Mendekar V. (25 February 2022), Machine Learning - it's all about assumptions [Online]. Available: https://www.kdnuggets.com/2021/02/machine-learning-assumptions.html [Accessed: March 11, 2022].
(30) Gumusbas E. (06 May 2020), First Trial with Random Forest [Online]. Available: https://github.com/ezgigm/Project4_Store_Sales_Forecasting/blob/master/STEP2_Random_Forest_Regressor.ipynb [Accessed: March 11, 2022].
(31) Komolafe A. (06 December 2021), Top Model Versioning Tools for Your ML Workflow [Online]. Available: https://neptune.ai/blog/top-model-versioning-tools\ [Accessed: March 11, 2022].
(32) ML Model Degradation, and why work only just starts when you reach production [Online]. Available: https://paulvanderlaken.com/2020/03/24/ml-model-performance-degradation-production-concept-drift/ [Accessed: March 11, 2022].
(33) Machiraju S. (01 November 2021), Why data drift detection is important and how do you automate it in 5 simple steps [Online]. Available: https://towardsdatascience.com/why-data-drift-detection-is-important-and-how-do-you-automate-it-in-5-simple-steps-96d611095d93 [Accessed: March 11, 2022].

(34) (14 February 2022), Artifacts [Online]. Available: https://www.jetbrains.com/help/idea/working-with-artifacts.html#artifact_configs [Accessed: March 11, 2022].

(35) Koundinya S. (25 September 2019), Walmart-recruiting-store-sales-forecasting [Online]. Available: https://medium.com/@sushma.koundinyam/walmart-recruiting-store-sales-forecasting-dc821e41f8be [Accessed: March 11, 2022].

(36) Brownlee B. (20 December 2017), A Gentle Introduction to Transfer Learning for Deep Learning [Online]. Available: https://machinelearningmastery.com/transfer-learning-for-deep-learning/ [Accessed: March 11, 2022].

(37) Chandan S. (02 October 2021), Explainable AI: Why Should Business Leaders Care? [Online]. Available: https://towardsdatascience.com/explainable-ai-why-should-business-leaders-care-5e5078c609b5 [Accessed: March 11, 2022].

(38) Watts J. (20 February 2022), How Responsible AI can equip businesses for success [Online]. Available: https://www.techradar.com/features/how-responsible-ai-can-equip-businesses-for-success [Accessed: March 11, 2022].

(39) Version Management [Online]. Available: https://its.unl.edu/bestpractices/version-management [Accessed: March 11, 2022].

(40) Hamilton L. (15 August 2020), Legal requirements for collecting personal data [Online]. Available: https://www.termsfeed.com/blog/legal-requirements-collect-personal-data/ [Accessed: March 11, 2022].

(41) (25 June 2021), How to improve data accuracy [Online]. Available: https://draycir.com/blog/how-to-improve-data-accuracy [Accessed: March 11, 2022].