Data Handling and Decision Making
Part 1: Essay

Contents

Task 1.1: Perform data gap analysis

• Briefly introduce the chosen organization or project.

The project aims to understand patient experiences with pharmaceuticals using machine learning, particularly NLP and sentiment analysis. **How can we gain comprehensive insights into the pharmaceutical landscape and patient experiences by analyzing drug reviews, temporal trends, review usefulness, and data distribution?**

• *Key Data Sources and Datasets Available:*

Drug Review Dataset:

Dataset includes patient reviews, drug names, conditions, and a 10-star rating system.

uniqueID: Unique identifier for each review entry.

drugName: Name of the drug being reviewed.

condition: Name of the medical condition for which the drug is prescribed.

review: Patient review detailing their experience with the drug.

rating: 10-star rating given by the patient.

date: Date of the review entry.

usefulCount: Number of users who found the review useful.

This dataset was collected from online pharmaceutical review platforms and featured in a study analyzing drug experiences, including effectiveness and side effects, through sentiment analysis.

-Data summary statistics of numerical columns:

	uniqueID	rating	usefulCount
count	161297.000000	161297.000000	161297.000000
mean	115923.585305	6.994377	28.004755
std	67004.445170	3.272329	36.403742
min	2.000000	1.000000	0.000000
25%	58063.000000	5.000000	6.000000
50%	115744.000000	8.000000	16.000000
75%	173776.000000	10.000000	36.000000
max	232291.000000	10.000000	1291.000000

- ## Inspect data integrity

-Data Cleaning:

Check for duplicates:
```
Duplicate Rows: 0
```

Check for missing values :
```
Missing Values:
uniqueID          0
drugName          0
condition       899
review            0
rating            0
date              0
usefulCount       0
dtype: int64
Duplicate Rows: 0
```

The dataset is 161297 rows × 7 columns, The missing values 899 are 0.5% of the total rows we have. So they have been drop. No duplicated values. Shape of Data Frame after removing missing values in the 'condition' column: (160398, 7).

• Identify gaps in data analytics and data protection.

Use statistical methods or machine learning algorithms to detect anomalies in the data, and the result are: [1 -1 1 ... -1 -1 -1]. Since the project aims to analyze drug reviews and ratings, anomalies detected by the Isolation Forest algorithm could potentially indicate unusual patterns or outliers in the data.

Objective	Goal	Gaps	Process or Test Name	Data Sources	Proposed Solutions
Objective 1	Analyze Drug Reviews and Ratings	- Lack of sentiment analysis for drug reviews - Absence of effectiveness metrics based on patient reviews	Sentiment Analysis Effectiveness Metrics	'drugName', 'condition', 'review', 'rating'	- Apply sentiment analysis techniques to extract sentiment from reviews - Develop metrics to measure the effectiveness of drugs based on patient reviews
Objective 2	Analyze Temporal Trends in Drug Reviews	- Lack of analysis on temporal trends in drug reviews - Absence of understanding of seasonal variations in review volume and drug ratings	Temporal Analysis	'date'	- Analyze changes in review volume over time - Identify seasonal variations in review volume and drug ratings - Explore trends in drug ratings over time
Objective 3	Explore Usefulness of Reviews	- Limited insight into how useful patients find reviews - Lack of understanding of patterns in review usefulness across drugs and conditions	Usefulness Analysis	'usefulCount'	- Calculate the average usefulness of reviews - Identify patterns in review usefulness across drugs and conditions
Objective 4	Assess Data Distribution and Labeling	- Uncertainty about the distribution of labels and counts - Lack of correlation analysis between labels/counts and other variables	Distribution and Correlation Analysis	'Label', 'Count'	- Conduct distribution analysis for labels and counts - Perform correlation analysis between labels/counts and other variables

Task 1.2: Recommend improvements

- **Reorganize current data-driven processes to enhance data analytics and decision-making.**

The following steps aim to reorganize current data-driven processes to enhance data analytics and decision-making:

1. **Define Objectives:**

Objective 1: Analyze Drug Reviews and Ratings

1. Goal: Understand the sentiment and effectiveness of different drugs based on patient reviews and ratings.
2. Key Questions:
    1. Which drugs have the highest and lowest ratings?
    2. What are the most common conditions for which drugs are prescribed?
    3. How do the ratings of drugs vary across different conditions?
3. Data Sources: 'drugName', 'condition', 'review', 'rating'

Objective 2: Analyze Temporal Trends in Drug Reviews

1. Goal: Identify temporal trends in drug reviews over time.
2. Key Questions:
    a. How has the volume of drug reviews changed over the years?
    b. Are there seasonal variations in the number of reviews?
    c. Are there any trends in the ratings of drugs over time?
3. Data Sources: 'date'

Objective 3: Explore Usefulness of Reviews

1. Goal: Investigate how useful patients find the reviews.
2. Key Questions:
    a. How many users found the reviews useful on average?
    b. Are there any patterns in the usefulness of reviews across different drugs or conditions?
3. Data Sources: 'usefulCount'

Objective 4: Assess Data Distribution and Labeling

1. Goal: Understand the distribution of labels and data ranges.
2. Key Questions:
    a. What are the distribution patterns for labels and counts?
    b. How are the labels and counts correlated with other variables?
3. Data Sources: 'Label', 'Count'

▪ Develop a roadmap for the development or enhancement of big data infrastructure.

1. Assess Current Data Infrastructure
    - Evaluate the Drug Review Dataset and its compatibility with analysis tools.
    - Identify any limitations or gaps in the dataset that may hinder analysis.
    - Generate a report outlining the strengths and weaknesses of the current data infrastructure.

2. Enhance Data Quality and Integrity
    - Conduct comprehensive data quality assessments for the Drug Review Dataset.
    - Implement data cleaning procedures to address any issues identified during the assessment.

3. Explore Advanced Analytics Techniques
    - Investigate advanced analytics models and algorithms for sentiment analysis and temporal trend analysis.
    - Test and refine these analytics models using the cleaned Drug Review Dataset.
    - Gather feedback from stakeholders to improve the effectiveness of the analytics techniques.

4. Foster Data-Driven Decision-Making Culture
- Conduct data literacy training programs for project team members to ensure they understand how to interpret and utilize the analyzed data effectively.
- Recognize and reward data-driven initiatives and successes to encourage a culture of data-driven decision-making.

5. Evaluation and Continuous Improvement
- Monitor the effectiveness of the implemented data analytics techniques in achieving project objectives.
- Collect feedback from stakeholders and end-users to identify areas for improvement.
- Continuously refine and optimize data analytics processes based on feedback and evolving project requirements.

▪ Address compliance aspects related to proposed changes in data analytics.

To enhance the compliance aspects of the proposed changes in data analytics, consider the following improvements:

Data Security and Privacy Measures: Implement robust data security protocols to safeguard patient information, including encryption for transmission and storage, access controls, and regular security audits. Comply with GDPR and HIPAA to protect patient privacy.

Ethical Considerations: Establish ethical guidelines for data collection, analysis, and usage to ensure responsible conduct. Provide ethics training for employees to raise awareness and promote ethical decision-making in data-related activities.

Transparency and Accountability: Maintain transparency by documenting data handling activities and communicating clearly with stakeholders about data usage. Establish accountability mechanisms to ensure compliance with regulations and ethical standards.

Continuous Compliance Monitoring: Implement regular compliance audits and assessments to monitor adherence to regulations and ethical guidelines. Designate a compliance team or officer to oversee compliance efforts and address issues promptly.

Task 1.3: Explain proposed big data analytics usage

Leveraging big data analytics can profoundly impact decision-making processes across various sectors. In healthcare, for instance, analyzing patient reviews and ratings can optimize treatment protocols, leading to improved patient outcomes and enhanced satisfaction. Additionally, identifying temporal trends in drug reviews enables healthcare providers to anticipate demand fluctuations and ensure timely availability of medications. Similarly, in the pharmaceutical industry, data analytics can drive insights into market trends, streamline drug development processes, and ensure compliance with regulatory standards, ultimately fostering innovation and competitiveness.
A compelling case study exemplifying the power of big data analytics is the utilization of sentiment analysis on patient reviews to enhance drug efficacy assessment. By systematically analyzing patient feedback and correlating it with clinical outcomes, pharmaceutical companies can identify areas for improvement, refine drug formulations, and tailor treatments to individual patient needs. This data-driven approach not only improves patient satisfaction but also enhances operational efficiency and drives innovation within the industry.

Overall, the strategic application of big data analytics enables organizations to make data-informed decisions, mitigate risks, and drive sustainable growth across diverse sectors.

References:

1. OneTrust. (2022, September 21). HIPAA vs. GDPR compliance: what's the difference? Retrieved from https://www.onetrust.com/blog/hipaa-vs-gdpr-compliance/
2. Cognilytica. (2024). Top 10 Ethical Considerations for AI Projects. Copyright © 2024. Retrieved from https://www.cognilytica.com/top-10-ethical-considerations-for-ai-projects/
3. Wren, H. (2024, January 18). What is AI transparency? A comprehensive guide. Retrieved from https://www.zendesk.de/blog/ai-transparency/#
4. Dilmegani, C. (2024, January 3). 5 Use Cases/Applications of Healthcare Sentiment Analysis in 2024. Retrieved from https://research.aimultiple.com/sentiment-analysis-healthcare/
5. Polisena, J., Andellini, M., Salerno, P., & Borsci, S. (2021, April). Case Studies on the Use of Sentiment Analysis to Assess the Effectiveness and Safety of Health Technologies: A Scoping Review. IEEE Access, PP(99), 1-1. DOI:10.1109/ACCESS.2021.3076356. Retrieved from https://www.researchgate.net/publication/351176672_Case_Studies_on_the_Use_of_Sentiment_Analysis_to_Assess_the_Effectiveness_and_Safety_of_Health_Technologies_A_Scoping_Review