AWS Cost Analysis for Amazon Best Selling Products

Final Report

Group 6:

Rimjhim Jain
Sudhanshu Dalvi
Saumit Chinchkhandi
Naval Gaikwad
Aniket Ghumed

Introduction

In the era of big data and advanced analytics, the ability to harness vast amounts of information for predictive insights has become crucial for businesses. This project, centered on the powerful capabilities of AWS cloud computing, leverages machine learning (ML) techniques to analyze and predict market trends, specifically focusing on identifying potential best-selling products on Amazon. The project is a testament to the synergy between cloud computing, data science, and e-commerce strategy.
The core objective of this endeavor is to deploy a machine learning model capable of sifting through extensive datasets to predict which products are likely to become best-sellers. This predictive analysis aims not only to demonstrate technical proficiency in handling large-scale data on AWS but also to provide actionable insights that could be pivotal for strategic business decisions in the e-commerce domain.

Project Scope
The project encompasses several critical aspects:

1. **Data Collection and Processing:**
    - Utilization of an expansive dataset from Amazon, comprising over 1.4 million products, enriched with various attributes like categories, ratings, prices, and reviews.
    - Meticulous data processing steps including merging, cleansing, and preparation, ensuring the data is primed for ML applications.
2. **Machine Learning Lifecycle:**
    - Employing AWS's robust cloud infrastructure to manage the machine learning lifecycle from data preparation to model training and evaluation.
    - Exploring various machine learning algorithms to identify the most effective approach for this specific use case.
3. **Predictive Analysis:**
    - Developing a model to predict the likelihood of products becoming best-sellers based on historical data trends.
    - Gaining insights into factors that drive product popularity and sales on Amazon.

Business Goals

The project is aligned with several key business goals:

- **Strategic Business Insights:** By predicting best-selling products, the project aids in understanding market trends, customer preferences, and potential future demands.
- **Optimized Inventory Management:** The insights garnered can help in optimizing stock levels, reducing overstock, and minimizing the risk of stockouts.
- **Targeted Marketing Efforts:** Understanding product trends enables more focused and effective marketing strategies, driving sales and customer engagement.
- **Data-Driven Decision Making:** The project exemplifies how data-driven insights can be pivotal in shaping business strategies in the e-commerce landscape.

Alignment with AWS Academy Cloud Foundations and Data Engineering

This project aligns with the educational program goals of AWS Academy Cloud Foundations and Data Engineering by:

- **Demonstrating Practical Application:** Applying theoretical knowledge of AWS cloud computing and data engineering in a real-world scenario.
- **Skill Enhancement:** Enhancing skills in cloud-based machine learning solutions, data manipulation, and analytical thinking.
- **Industry Relevance:** Addressing a current and relevant problem in e-commerce, making the project both educational and pertinent to industry needs.

Data Preparation

The foundation of any successful machine learning project lies in the quality and preparation of the data. For this project, we focused on an extensive dataset from Amazon, providing a rich, multifaceted view into the e-commerce giant's product landscape.

1. **Data Collection:**
    - Sourcing of data from Amazon, encompassing a diverse array of product categories, customer reviews, pricing information, and sales metrics.
    - The dataset includes two primary CSV files: 'amazon_categories.csv' and 'amazon_products.csv', each offering unique insights into product classifications and individual product details.

2. **Data Integration:**
    - Merging of 'amazon_categories.csv' and 'amazon_products.csv' to create a unified dataset. This process involved aligning common keys and ensuring that the integration did not lead to data redundancy or loss.
    - Validation of the merged dataset to guarantee the integrity and consistency of data across different categories and product attributes.

3. **Data Cleaning:**
    - Identification and handling of missing values to prevent inaccuracies in the analysis. This step was crucial in ensuring that the machine learning models were trained on complete and reliable data.
    - Standardization of data formats, particularly in numerical and categorical variables, to facilitate smooth processing and analysis.

4. **Initial Exploration:**
    - Examination of the first few rows of the data to gain a preliminary understanding of the dataset's structure and the nature of the variables involved.
    - Utilization of summary statistics to explore the central tendencies, dispersion, and overall distribution of the key numerical variables.

Initial Analysis

The initial analysis phase aimed to extract foundational insights from the data, setting the stage for more complex machine learning algorithms.

1. **Statistical Overview:**
    - Generation of descriptive statistics to summarize central tendencies, dispersion, and shape of the dataset's distribution, especially for variables like product prices and customer ratings.
    - Evaluation of data types and unique values in each column to understand the dataset's composition and identify any anomalies or irregularities.

2. **Unique Value Analysis:**
    - Assessment of unique values across different columns to determine the diversity and range of data. This step was vital for understanding the granularity and depth of the dataset.
    - Identification of potential categorical variables that could be leveraged for segmenting the data in later analysis stages.
3. **Missing Data Analysis:**
    - Quantification of missing values, particularly in critical columns like 'id' and 'category_name', to assess the extent of data completeness.
    - Strategic handling of missing data, either through imputation or exclusion, based on their impact on the overall dataset and subsequent analyses.
4. **Data Visualization:**
    - Preliminary visualization using tools like Matplotlib, Seaborn, and Plotly to observe patterns, trends, and outliers in the data.
    - Creation of plots like histograms, scatter plots, and bar charts to visually explore relationships between key variables.

This phase of data preparation and initial analysis was instrumental in shaping the direction of the machine learning project. By ensuring the data was clean, integrated, and well-understood, we laid a solid foundation for building predictive models that are both accurate and reliable. The insights gained here also provided a glimpse into the complex dynamics of Amazon's product ecosystem, setting the stage for deeper analysis and predictive modeling.

Data Visualization

Data visualization is a powerful tool for understanding complex datasets. It enables the identification of patterns, trends, outliers, and relationships within the data. In this project, we employed various visualization techniques to gain deeper insights into Amazon's product data.

1. **Category Analysis:**
    - Utilized bar charts to display the distribution of products across different categories, highlighting the most prevalent categories on Amazon.
    - Analyzed the average price and rating by category using grouped bar charts, which provided insights into customer preferences and pricing strategies across different segments.

2. **Best Selling Products Analysis:**
    - Developed visualizations to compare the characteristics of best-selling products against others. This included scatter plots to correlate prices with ratings, highlighting how these factors influence the best-seller status.
    - Identified the top best-selling products using bar charts, focusing on aspects like customer ratings, number of reviews, and pricing.

3. **Price Distribution Analysis:**
    - Created histograms and box plots to examine the distribution of product prices. This helped in identifying the range of prices across the dataset and pinpointing any outliers.
    - Investigated the presence of any high-end products or significantly low-priced items, which could indicate special categories or promotional strategies.

4. **Ratings and Reviews Analysis:**
    - Employed scatter plots to explore the relationship between customer ratings and the number of reviews. This analysis aimed to understand if highly rated products generally receive more reviews.
    - Analyzed the average rating and review count for products to gauge overall customer satisfaction and engagement.

Analysis

The visualizations facilitated a comprehensive analysis of the dataset, leading to several key insights:

1. **Category Insights:**
    - Certain categories dominated the dataset, suggesting higher customer demand or larger product availability in these segments.
    - Average prices and ratings varied significantly across categories, indicating different customer expectations and market dynamics in each segment.

2. **Best Sellers Insights:**
    - Best-selling products exhibited specific trends in pricing and ratings, underscoring the importance of these factors in driving sales.
    - The analysis of best sellers provided a blueprint for what makes a product successful on Amazon, which could guide future inventory and marketing strategies.

3. **Price Trends:**
    - The price distribution analysis highlighted the diversity of products on Amazon, ranging from budget to premium.
    - Understanding price trends helped in identifying potential market gaps and opportunities for new product introductions.

4. **Customer Feedback and Engagement:**
    - The relationship between ratings and reviews offered insights into customer engagement. Products with higher ratings tended to attract more reviews, reinforcing their popularity.
    - The average ratings and review counts helped in assessing the overall customer sentiment and satisfaction with the products.

The data visualization and analysis phase was crucial in extracting meaningful insights from the Amazon dataset. It not only facilitated a better understanding of the current e-commerce landscape but also provided valuable inputs for the subsequent machine learning model development. These insights were instrumental in predicting which products have the potential to become best-sellers, ultimately aiding in making informed business decisions.

Predictive Modeling

Overview

Predictive modeling forms the core of this project, leveraging Amazon's extensive product data to forecast future trends, specifically predicting potential best-sellers. The approach involved using Amazon SageMaker, a robust and scalable machine learning service provided by AWS.

Model Setup and Training

1. **SageMaker Estimator Creation:**
    - Initialized a SageMaker SKLearn Estimator, a managed model training service that simplifies the machine learning workflow.
    - Configured the estimator with necessary parameters including the entry point script ('train.py'), the source directory containing training scripts and requirements, the IAM role, and instance specifications.

2. **Training Job Execution:**
    - The model training process was initiated by pointing to the Amazon S3 bucket where the training data resides.
    - SageMaker's streamlined process allowed for efficient model training without the need to manage underlying resources manually.

Model Training Script

The train.py script in the directory encompassed the following key components:

1. **Data Preprocessing:**
    - Included necessary steps to clean, transform, and prepare the Amazon product data for modeling.
    - Techniques like handling missing values, encoding categorical variables, and feature scaling were employed to optimize the data for machine learning algorithms.

2. **Model Selection and Training:**
    - Chose appropriate machine learning algorithms considering the nature of the data and the prediction objective. Logistic regression was a primary choice due to its efficacy in binary classification problems.
    - Implemented model training code that allowed the algorithm to learn from the historical data, focusing on identifying patterns that lead to a product becoming a best-seller.

3. **Hyperparameter Tuning:**

    - SageMaker's hyperparameter tuning capability was utilized to automatically find the optimal model parameters, enhancing model performance.
    - This process involved running multiple training jobs with different hyperparameter combinations to determine the most effective settings.

Deployment and Prediction

After training, the model was deployed to a SageMaker endpoint for real-time predictions. The deployed model could then be used to predict the likelihood of products becoming best-sellers based on their features like customer reviews, ratings, and pricing.

Challenges and Solutions

During the deployment phase, a common challenge faced was ensuring the SageMaker endpoint was correctly configured and responsive. This included:

- Proper implementation of model loading (`
model_fn) and prediction (predict_fn) functions in the train.py` script.

- Addressing any compatibility issues between the training environment and the prediction endpoint, such as library versions and data formats.

The predictive modeling phase of this project successfully leveraged AWS SageMaker's capabilities to train and deploy a machine learning model. The model's predictions can be instrumental in guiding business strategies, inventory planning, and marketing efforts on Amazon's platform. By analyzing product features and customer feedback, the model offers valuable predictions about potential best-sellers, enabling data-driven decision-making in the dynamic e-commerce landscape.

Conclusion:

This project stands as a testament to the transformative power of cloud computing and machine learning in the modern business landscape. It not only provided valuable insights and predictions but also set a precedent for future projects that aim to harness the power of AWS and machine learning to drive innovation and business success. The methodologies and lessons learned from this project can be applied across various domains, showcasing the versatility and impact of cloud-based machine learning solutions.