Public Cloud Marketplace Competitive Analysis and Sales Prediction

CISC7201 INTRODUCTION TO DATA SCIENCE PROGRAMMING

Group AZ

December 16, 2024

1 Introduction and Motivation

In the domestic market, leading cloud service providers such as Alibaba Cloud, Tencent Cloud, and Huawei Cloud have developed extensive marketplaces that offer a diverse array of enterprise-level solutions, including infrastructure, security, artificial intelligence, and the Internet of Things.

This project evaluates the competitive strengths of major cloud service providers by analyzing product types, pricing strategies, delivery methods, and other critical factors. Additionally, it seeks to forecast sales based on these features.

2 Research Significance

The exponential growth of cloud computing has revolutionized the IT industry, enabling businesses to scale operations efficiently and drive innovation. Understanding the competitive dynamics within the cloud market is crucial for organizations to make strategic decisions when selecting cloud service providers.

This analysis provides a comprehensive overview of the current state and future trends of the public cloud marketplace. Additionally, it offers actionable insights for businesses to optimize their cloud investments effectively.

3 Data Collection and Preprocessing

Identifying target cloud service providers, verifying web scraping permissions, analyzing web page structures, and developing scripts to collect and preprocess data.

3.1 Data Source

The data was selected the following major cloud service providers in China:

1. **Alibaba Cloud:** Asia's largest cloud platform, founded in 2009.
2. **Tencent Cloud:** Launched in 2013, holds 18% of the domestic market.
3. **Huawei Cloud:** Established in 2011, with an 8% market share.
4. **eSurfing Cloud:** China Telecom's brand, registered in 2016.
5. **QingCloud:** Independent provider, founded in 2012.
6. **Kingsoft Cloud:** Part of Kingsoft Corporation, established in 2012.

3.2 Data Collection

The data was obtained from the official websites of the aforementioned six cloud service providers. Prior to the data crawling, the compliance should be checked through the robots.txt files of the websites to determine the accessible and inaccessible parts. It could be found that Alibaba Cloud, Tencent Cloud, and Huawei Cloud had relatively relaxed crawler access restrictions, and their set crawling boundaries allowed for data collection within a larger scope.

However, the robots.txt files for QingCloud, eSurfing Cloud, and Kingsoft Cloud did not be found. Before initiating the formal crawling task, the targeted test scripts were wrote and confirmed that the target data for these cloud markets could be successfully crawled without triggering any violations or encountering any abnormal situations. This allowed to proceed with the subsequent large-scale and systematic crawling work.

3.3 Data Preprocessing

After data crawling, the data types, missing values, and outliers were checked.

1. **Alibaba Cloud:** Crawled 5,001 rows and 13 columns. Converted text-based numeric columns to numeric and date columns to date format. Removed entirely missing Launch Time column. Handled missing values by replacing Rating Count with 0 and creating a ”Product Score Missing” column. Filled 29 missing Price values with 0.1. Removed anomalous values in Product Score, Response Time, and SLA.

2. **Tencent Cloud:** Crawled 8,717 rows and 7 columns. No data type transformations required. Filled missing values based on domain expertise.

3. **Huawei Cloud:** Crawled 9,425 rows and 12 columns. Converted text-based numeric columns to numeric. Removed Success column and rows with ”n/a” values. Manually filled the sole missing value in Delivery Type.

4. **QingCloud:** Crawled 382 rows and 7 columns. Converted Publish Time column to standardized date format. Manually filled minimal missing values.

5. **eSurfing Cloud:** Crawled 117 rows and 7 columns. Converted Price field to numeric. Deleted rows with missing data due to limited number and inability to find corresponding product names.

6. **Kingsoft Cloud:** Crawled 245 rows and 11 columns. No data type transformations required. Filled missing values based on domain expertise.

4 Exploratory Data Analysis and Visualization

Examining product category versatility, delivery methods, and pricing strategies among leading and emerging cloud service providers in China.

4.1 Competitive Analysis

4.1.1 Product Category Analysis

Process: Began by loading datasets from Alibaba, Huawei, and Tencent, focusing on key columns such as product name, description, and category, the work about defining product categories and identifying relevant keywords had been done. Products were classified based on the presence of these keywords in titles and descriptions, with unmatched products labeled as ”Other.” To improve classification accuracy, the analysis of high-frequency words within the ”Other” category had been done. Finally, the results were visualized by using grouped bar charts.

Conclusion: Common demands across providers include operations and management, security services, and authentication and verification. However, there are notable differences: Alibaba Cloud emphasizes infrastructure and enterprise solutions, with less focus on small-scale e-commerce or mini-programs. Huawei Cloud prioritizes security, compliance, and authentication, catering to enterprise-level customers with a strong focus on data protection. Tencent Cloud specializes in mini programs and e-commerce, leveraging its integration with Tencent’s WeChat ecosystem to offer robust e-commerce and lightweight development tools.

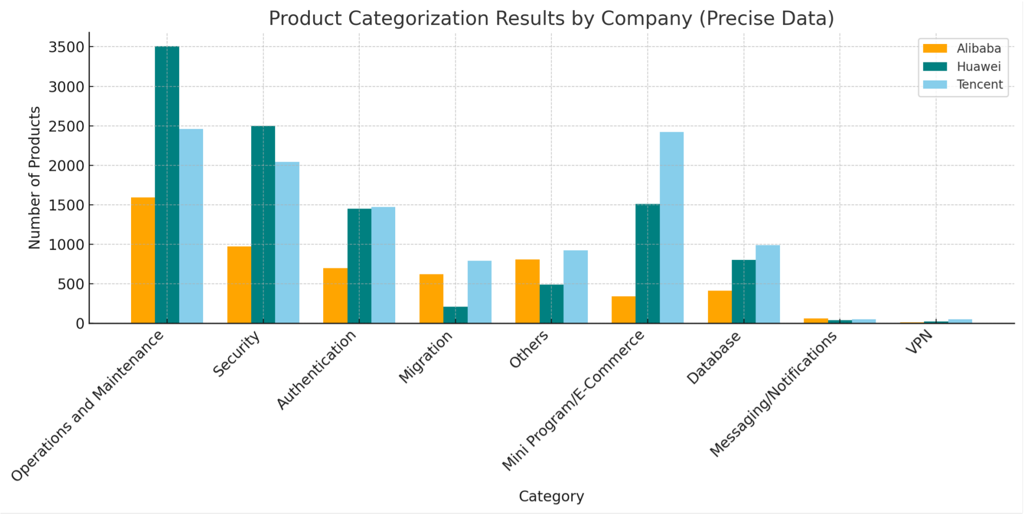


Figure 1: Product Category Distribution

4.1.2 Delivery Type Analysis

Process: It began by loading datasets from Alibaba, Huawei, and Tencent, verifying the delivery method field, and standardizing the field names. The main delivery method categories (API, SaaS, Manual Service, Image) were then defined and be classified any unmatched methods as "Other." Rules were applied to classify the records and calculate the proportions for each category. Creating pie charts for each company to display the distribution of delivery methods was the final step, followed by combining them into one comparative chart.

Conclusion: The analysis reveals distinct delivery method preferences among the providers. Alibaba predominantly uses API-based delivery methods (65.8%), reflecting its emphasis on development interfaces and platform services. It also offers Manual Service (15.5%) and SaaS (12.9%), showcasing a diverse service range. Huawei primarily relies on licenses (67.2%), focusing on software and service licenses within its cloud ecosystem, with smaller shares in Manual Service (10.9%) and Hardware (11.1%). Tencent's main delivery method is Manual Service (56.1%), underscoring its strength in customized solutions, while SaaS (22.1%), API (13.4%), and Image (7.6%) indicate a need for more technical and standardized services. In summary, Alibaba and Huawei prioritize technical and standardized services (API, SaaS, licenses), whereas Tencent excels in manual, customized services.

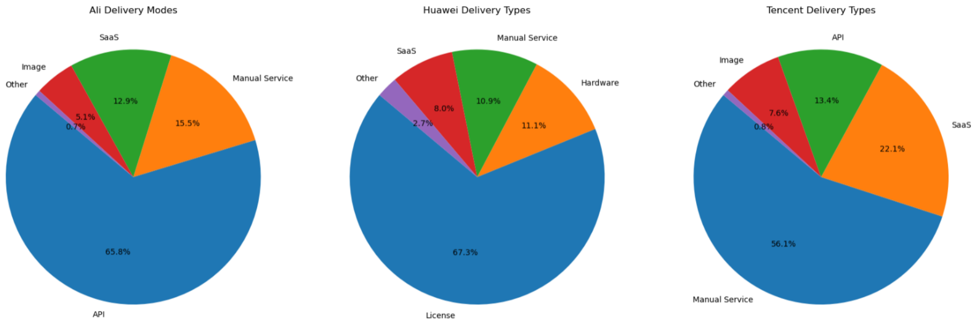


Figure 2: Delivery Type Distribution

4.2 Price Comparison

Process: The price distribution was compared between leading (Alibaba, Tencent, Huawei) and emerging (eSurfing Cloud, QingCloud, Kingsoft Cloud) cloud providers. First step was to extracted and cleaned the price data, converting it to numeric format. Next, the providers were categorized into two groups: Leading and Emerging. It also removed outliers (values outside the 95th percentile) for clear visualization. Finally, the box plots were used to visualize the price distribution.

Conclusion: The median price for leading providers was significantly lower than for emerging providers, indicating more uniform pricing among the leading providers. Emerging providers showed greater price variability and higher ranges, suggesting a focus on high-end, customized products. The upper quartile of emerging providers' prices could exceed 4,000 units, while leading providers' prices remain lower, reflecting standardized, large-scale services. These differences highlight distinct market strategies: leading providers target enterprises and small businesses with competitive pricing, while emerging providers offer bespoke, high-value services.

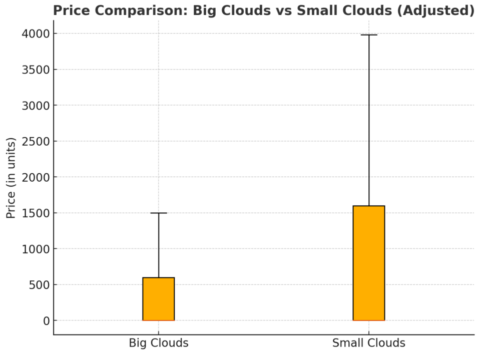


Figure 3: Product Price Distribution

5 Model Training and Sales Forecast

Processing and cleaning data, training models, and predicting sales to provide actionable insights for optimizing cloud investments.

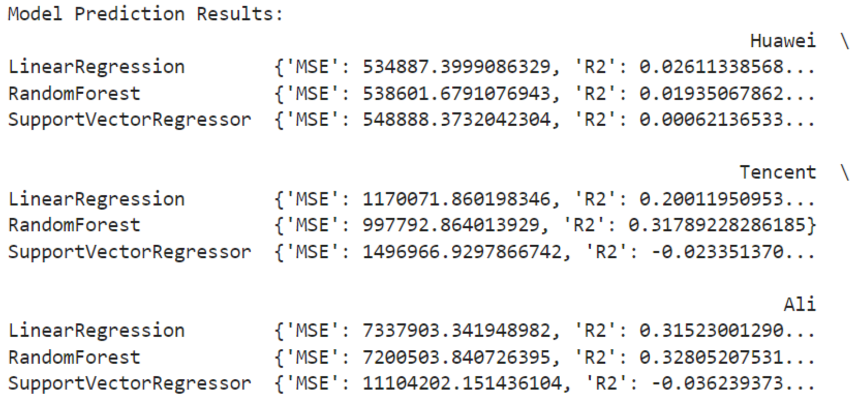
5.1 Data Processing and Cleaning

1. **Huawei Data:**
    - Handled missing values: Filled in Product Score, Price Unit; removed records with missing Delivery Type.
    - Cleaned up redundant information: removed empty strings in the Categories column.
    - Standardized types: Ensured consistent format for Price Number and Price Unit.
2. **Tencent Data:**
    - Handled missing values: Deleted missing records of Title and Delivery Type.
    - Standardized types: Converted Price to floating point.
3. **Alibaba Data:**

- Handled missing values: Filled in Product Score and extracted Transaction Count and Rating Score values.
- Cleaned invalid columns: Deleted the empty column Launch Time.
- Standardized types: Extracted and standardized Price values.

5.2 Model Training and Sales Forecast

The `train_test_split` was used to split the data into a training set 80% and a test set 20%. The model was trained exclusively on the training set, ensuring that the test set remained untouched during training. After completing the model training, the test set was used for predictions and compared the predicted results with the actual values from the test set. Then the calculated evaluation metrics such as Mean Squared Error (MSE) and R^2 could be gotten to assess the model's performance.



```
Model Prediction Results:
                                                                  Huawei  \
LinearRegression       {'MSE': 534887.3999086329, 'R2': 0.02611338568...
RandomForest           {'MSE': 538601.6791076943, 'R2': 0.01935067862...
SupportVectorRegressor {'MSE': 548888.3732042304, 'R2': 0.00062136533...

                                                                 Tencent  \
LinearRegression       {'MSE': 1170071.860198346, 'R2': 0.20011950953...
RandomForest           {'MSE': 997792.864013929, 'R2': 0.31789228286185}
SupportVectorRegressor {'MSE': 1496966.9297866742, 'R2': -0.023351370...

                                                                     Ali
LinearRegression       {'MSE': 7337903.341948982, 'R2': 0.31523001290...
RandomForest           {'MSE': 7200503.840726395, 'R2': 0.32805207531...
SupportVectorRegressor {'MSE': 11104202.151436104, 'R2': -0.036239373...
```

Figure 4: Model Prediction Results

The following is a detailed analysis of the forecast results:

1. **Huawei Data**
   - Linear regression (R^2: 0.026): The explanatory power is weak, explaining only 2.6% of the sales fluctuation.
   - Random Forest (R^2: 0.019): Slightly lower than linear regression.
   - Support Vector Regression (R^2: 0.001): Almost no explanatory power.
   - Summary: Sales volume is affected by nonlinear or complex factors, and data quality and insufficient feature engineering are the main problems.
2. **Tencent Data**
   - Linear regression (R^2: 0.20): fair explanatory power.
   - Random Forest (R^2: 0.32): performs best and captures non-linear relationships.
   - Support Vector Regression (R^2: -0.02): Poor performance.
   - Summary: Further exploration of potential features such as user behavior or company size is needed.
3. **Alibaba Data**
   - Linear regression (R^2: 0.32): Fair performance.

- Random Forest (R^2: 0.33): Slightly better than linear regression and captures complex patterns.
- Support Vector Regression (R^2: -0.03): Worst performance.
- Summary: Random Forest performs well, and features such as price and rating have nonlinear relationships.

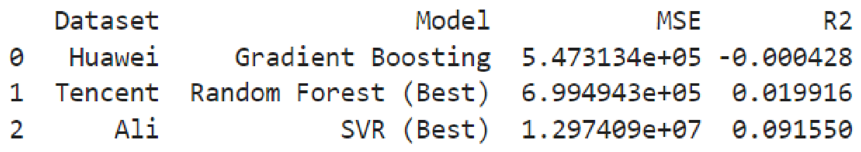
5.3 Further Optimizations

5.3.1 First Round of Improvements

The first round of improvements include:

1. **Feature Engineering**
   - Data cleaning and standardization.
   - Basic features such as ratings, number of reviews, etc. are extracted.
2. **Model optimization**
   - The gradient boosting model was used to optimize the Huawei dataset.
   - random forest and support vector machine are adjusted through grid search (GridSearchCV) to improve the prediction ability of Tencent and Alibaba datasets.

Here are the results of trying to optimize the predicted sales:



```
   Dataset                 Model           MSE        R2
0   Huawei     Gradient Boosting  5.473134e+05 -0.000428
1  Tencent  Random Forest (Best)  6.994943e+05  0.019916
2      Ali            SVR (Best)  1.297409e+07  0.091550
```

Figure 5: First Round of Improvements

Summary of the improved results:

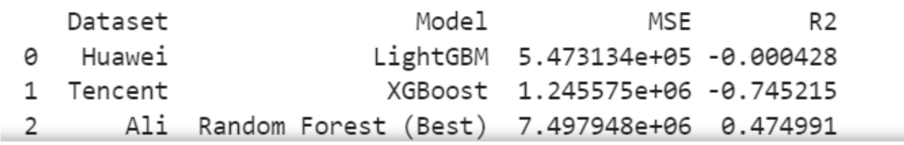
1. **Huawei Data:** The performance is still poor and needs to be further optimized through feature engineering and outlier processing.
2. **Tencent Data:** There is a slight improvement, but the R^2 is low, and the feature interpretation ability still needs to be enhanced.
3. **Alibaba Data:** SVR performs worse than linear regression, which could try other nonlinear models (such as XGBoost).

5.3.2 Second Round of Improvements

In response to the problem of the previous improvement, make new improvements below. However, the prediction results are still not ideal.

The current problems may be caused by the following points:

1. **Insufficient data features**



```
   Dataset                Model           MSE        R2
0   Huawei             LightGBM  5.473134e+05 -0.000428
1  Tencent              XGBoost  1.245575e+06 -0.745215
2      Ali  Random Forest (Best)  7.497948e+06  0.474991
```

Figure 6: Second Round of Improvements

- The features used for prediction in the data may not fully represent the target variable (sales). For example, the existing features may have little impact on the change in sales, or may lack key influencing factors (such as seasonality, promotional activities, market competition, etc.).

2. **Data quality issues**

   - Noise, outliers, or missing values in the data are not handled effectively.
   - Categorical features are not properly encoded, such as Categories and Merchant features, which have a direct impact on sales.

3. **Model selection problems**

   - LightGBM, XGBoost, and Random Forest models are used. These models may overfit or underfit under certain conditions, especially when the amount of data is small or the features are insufficient.

6 Conclusion

This project provides an exploratory analysis of the competitive landscape among major cloud service providers in China. By examining product versatility, delivery methods, and pricing strategies, it offers valuable insights into the strengths and market positioning of these providers. Despite challenges in data quality and feature engineering, the findings highlight the distinct approaches of leading and emerging providers, aiding businesses in making informed cloud investment decisions.

Future work will focus on enhancing data features and exploring advanced models to improve the accuracy of sales forecasting.