Data Analysis for Mining Sector and Petrographic Lab

**Introduction and discovery:**

Petrographics lab is a leading provider of petrographic sample prep, analysis, and consulting services in the region. Since I am working part-time for the Petrographic Lab. I would like to perform data analysis for the company since it would be benecifial for me and and company.

The laboratory holds a pivotal position in supporting research studies in universities and various industries, including mining, petroleum, and construction. With an emphasis on geological sample preparation and petrographic analysis, the laboratory's services are vital for advancing scientific understanding and facilitating decision-making in these sectors.

Problems that we try to address:

* Can we predict the political situation of countries based on certain features to guide investment decisions for mining operations? Knowing the political situation in a region can help in making informed decisions about mining investments, considering stability and potential risks.
  + Certain features, such as political indicators or economic conditions, may influence mining investments and can be used to predict the political situation.
* Can we determine an optimal pricing strategy based on quantity or other variables to maximize profit? Understanding the relationship between pricing, other variables, and profit can help in setting competitive prices and maintaining profitability.
  + I assume that there is a correlation between quantity or other variables and profit, suggesting that higher quantity orders may lead to higher profits.
* What are the different customer segments based on their order completion time and ratings? Can we identify patterns in the completion time of orders and customer ratings?
  + Customers with shorter order completion times tend to give higher ratings, indicating higher satisfaction levels.

**Dataset Preparation:**

Two different dataset is used for this project:

1 - Dataset that provides information about the total world production of mineral raw materials

<https://www.world-mining-data.info/?World_Mining_Data___Data_Section>

2 - Dataset from their database("VanPetroDatabase"). Since they know, I am going to submit it to college, they manipulated the data and if they like what I am trying to do, they will let me do analysis with their real data in future.

**Dataset 1**

This dataset will serve as resources for understanding world mining, making informed decisions, and drawing meaningful conclusions about the mining industry on global scale.

* PetroDataset.xlsx:

This dataset provides information about the total world production of mineral raw materials for a specific year.

* GroupsCommodities.xlsx:

This dataset presents data on the production of mineral raw materials categorized by the specific type of mineral.

* IncomeLvl.xlsx:

This dataset includes data on the annual per capita income of different countries involved in the mining industry.

* Political stability.xlsx:

This dataset offers information about the political stability of various countries or regions where mining activities occur.

My aim is to combine different data sets to provide data analysis about mining sector.

Info about final cleaned dataset and features:

- year (continuous): The year in which the order was collected or recorded.

- value (continuous): A numerical value associated with volume of material production.

- income\_group\_encoded (categorical): An encoded representation of the income group to which the sample belongs.

- groupscommodity\_encoded (categorical): An encoded representation of the commodity group.

- region\_encoded (categorical): An encoded representation of the region.

- politicsest\_encoded (categorical): An encoded representation of the political situation.

**Dataset 2**

Dataset from their database("VanPetroDatabase.xlsx").

Info about dataset and features:

- Quantity: Quantity of material.

- TotalWeight: Total weight of the material.

- OrderCompleteHour: Total hour duration at which the order was completed.

- Invoice\_number: Invoice number associated with the order

- Profit: Profit earned from operation.

- Price: Price of service given to customers.

- Rating: Rating given by customers about their satisfactory level.

- Customer.companyName: Name of the customer's company.

- Customer.Country: Country where the customer is located.

- Service.Service: Name of the lab service.

- DelayIssues.DelayType: Type of delay issue.

- Order.Material/RockType: Type of material or rock extracted in the order.

- Order.CompletionDate: Date of order completion.

The top rows of the dataset can be viewed in the figure below for Dataset 1. A screenshot of a computer

Description automatically generated

Figure 1. ‘Dataset 1 head()’

**Data cleaning and transformation for Dataset 1**

The original dataset had separate columns for each year (e.g., 2017, 2018, 2019, etc.). These columns were melted into a single 'Year' column. The original dataset was merged with two additional datasets ('Politicalstability.xlsx' and 'IncomeLvl.xlsx') using the 'Country' column as the primary key. Null values in the 'Value' column and 'Lending category' column were handled by dropping.

The 'PoliticsEst' column was classified into 'Stable' and 'Unstable' categories based on a threshold value of 0. If 'PoliticsEst' was less than 0, it was considered 'Unstable', otherwise 'Stable'. Categorical columns ('Income group', 'GroupsCommodity', 'Region', and 'PoliticsEst') were labeled encoded to convert them into numeric values. The 'Value' column was log transformed to address data scale the values.

'Year' column was converted numeric data type. Column names were modified to be in lowercase.

The top rows of the dataset for Dataset 2. A screenshot of a computer

Description automatically generated

Figure 2. ‘Dataset 2 head()’

**Data cleaning and transformation for Dataset 2**

Removed 'DelayIssues.DelayType' column due to many missing values. Extracted the last two digits from the 'Order.SubmissionDate' and 'Order.CompletionDate' columns and converted them to integers to represent the year. Replaced dots in column names with underscores for better compatibility.

Converted the 'customer\_isbigaccount' column from bool to integer (0 and 1) to represent the True and False values. Transformed the 'customer\_companyname' column by grouping certain companies under the 'Others' category. Dropped unnecessary columns that were not useful for analysis. Created dummy variables for the categorical columns 'companyname', 'servicetype', and 'materialtype' using one-hot encoding to convert them into numeric values.

**Model Implementation**

Based on the World Mining dataset 1, task is to predict the "politicsest\_encoded" variable based on the other independent variables. The best model for this classification task is the Support Vector Machine (SVM) with an RBF kernel, C=10, and gamma=1.

Justification for the SVM Model:

SVM is a powerful classification algorithm suitable that can handle complex relationships between the independent and dependent variables. The dataset contains a mix of numerical and categorical features, and SVM handles numerical features effectively with dummy variables. The best SVM model achieved an accuracy of around 88%, which is reasonable for this classification task.

Objective is to predict the "politicsest\_encoded" variable, which represents the political situation of countries. The SVM model helps achieve this objective by classifying countries into two political classes based on the other independent variables.

Hypotheses Testing:

The SVM model allows us to test the hypothesis that certain independent variables, such as income level, region, material type, and volume of commodities. The model's performance metrics, such as precision, recall, and accuracy, help in assessing how well the model predicts the political situation. The confusion matrix provides insights into the true positive, true negative which can help to understanding the model's strengths and weaknesses.

Based on "VanPetroDatabase" Dataset 2, the proposed machine learning model is a Ridge Regression model with a specific alpha value that has the highest R2 score and lowest RMSE.

The chosen techniques facilitate testing of the hypotheses and provide insight into the modeling objectives as follows:

* Feature Selection: The analysis used three different feature selection techniques: Correlation-based selection, Variance threshold selection, and SelectKBest.
* Feature Transformation: Polynomial feature transformation (Poly 2 Interaction) was applied to selected features.
* The analysis used R2 and RMSE scores to evaluate the performance of different models.
* Several regression techniques are considered, including linear regression and Ridge Regression with the chosen alpha value demonstrated the best performance.

Clustering analysis is applied again for Dataset 2 "VanPetroDatabase", the model is the KMeans clustering algorithm with 5 clusters. KMeans is that it performed better in terms of clustering quality, as evidenced by higher Silhouette Score and Calinski Harabasz Score, and a lower Davies Bouldin Score compared to the Agglomerative method. The clustering analysis helps group similar data points together based on the features "ordercompletehour" and "rating." By clustering the data, we gain insights into customer behavior and preferences.

**Results Interpretation and Implications**

1. Classification model result

A graph with numbers and a few words

Description automatically generatedA screenshot of a computer screen

Description automatically generated

Figure 3. ‘Confusion Matrix and Classification Report’

The SVM classification model appears to be accurate on the test data. The accuracy of the test set is approximately 0.86 which means that the model correctly predicts political situation of countries with an 86% accuracy rate.

* The parameter values chosen for the SVM model have been determined through grid search and cross-validation, optimizing the hyperparameters for the given dataset. The chosen values (C=10 and gamma=1) are based on the best performance during the grid search process
* The model's accuracy rate of 86% is quite good and suggests that the model can make accurate predictions most of the time.
* Our company or others can predict beforehand the political situation before making massive invesment since political situation will affect our bussines for invested country.

The analysis used SVM classification to predict the political situation based on various features, including income level, region, material type, and volume of commodities. Insights derived from the analysis can be used by government agencies and businesses to gain a better understanding of the factors that influence political situations in different countries.

1. Regression model result

The obtained results are presented in Figure 4, which includes a summary table of different feature selection combinations and a scatter plot depicting the predicted values versus the actual values.

A screenshot of a computer screen

Description automatically generated A graph with red dots

Description automatically generated

Figure 4. Summary Analysis and Predicted vs Actual Value scatter plot.

The selected Ridge Alpha regression model has R2 value of 0.905582, indicating that approximately 90.6% of the variance in the target variable (profit) can be explained by the model.

The model output and behavior seem reasonable. The coefficients of the features provide insights into their impact on profit. For instance, higher total weight has a positive impact on profit.

While no model is perfect, the selected model with Ridge regularization performs well in terms of R2 and RMSE, minimizing errors in profit predictions.

Key Findings and Insights:

* The Ridge Alpha regression model with feature transformation and scaling has been shown to be the best performing model based on the R2 and RMSE scores.
* Feature selection techniques, such as correlation-based selection, variance threshold, and SelectKBest, were used to identify relevant features for the model.

Our company gives price to customers according to quantity, but our analysis shows that profit and totalweight has higher correlation than quantity, so suggestion will be given to charge their service to customers as per totalweight i/o quantity.

1. Clustering model result

A screenshot of a phone

Description automatically generated

A screen shot of a computer generated image

Description automatically generated

Silhouette, Calinski Harabasz, and Davies Bouldin Score used to evaluate the clustering performance. A higher Silhouette Score and Calinski Harabasz Score indicate better clustering, while a lower Davies Bouldin Score suggests more separate clusters.

* Based on the evaluation metrics, the KMeans clustering model appears to have produced reasonable clustering results, with a Silhouette Score of around 0.56.
* 5 clusters are used. It is used the Elbow method and KneeLocator to determine the optimal number of clusters.
* The goal is to segment customers based on their rating and order completion hour, the model's performance.
* KMeans in this case is reasonable given the dataset and goals. However, other clustering algorithms, such as DBSCAN or hierarchical clustering, could also be explored if they might better suit the problem.
* For cluster 1, 2, 4 seems gather around rating 0 values no matter when order is completed in hour. It is weird because we expect if we complete orders earlier, the rating would be higher. After discussing with the company, they doubt customers do not pay attention to the rating process after completing their order. Also, if customers do not give rating, the system assigns a default value that is why most of data spread in specific value.

**Out-of-sample Predictions:**

Synthetic dataset is generated with named synthetic to simulate out-of-sample predictions for our classification and regression model. The synthetic dataset was created by randomly selecting 50 rows from the original DataFrame, and then for each random sample, we calculated the 25th and 50th percentiles.

The classification model achieved impressive performance on the test dataset, with an accuracy of 0.9412. The best hyperparameters for the SVM model were determined to be C=1, gamma=1, and the kernel used was 'rbf'. These parameters seem to be well-suited for the given classification task.

Classification report reveals that the model struggles with classifying samples of "class 1," as indicated by the lower precision (0.50) and f1-score (0.67) for this class. While the recall (1.00) is high for "class 1," it is important to recognize that this is based on a small number of samples (support=4).

The regression model result does not differ much from the last step. Based on scatter plot, strong correlation observed between the predicted and actual values, it appears that the Ridge regression model performs well on the test data (out-of-sample data). The model demonstrates strong capabilities, as it can make accurate predictions on new data.

**Concluding Remarks**

In this data analysis project for the mining sector and Petrographic Lab, we performed data analysis on two datasets - "Dataset 1" providing information about the total world production of mineral raw materials, and "Dataset 2" from the Petrographic Lab's database, containing details about customer orders and mining operations. We addressed several key business questions and implemented machine learning models to predict political situations, determine pricing strategies, and perform customer segmentation.

Major Findings:

* Classification Model: The SVM classification model was developed to predict the political situation of countries based on various features, such as income level, region, material type, and commodity volume. The model achieved an accuracy rate of approximately 86% on the test data, indicating its ability to accurately classify countries into 'Stable' or 'Unstable'.
* Regression Model: For optimizing pricing strategies and maximizing profit, we implemented a Ridge Regression model. The model achieved an R2 score of 0.906, indicating that around 90.6% of the variance in profit can be explained by the selected features. The model suggests that total weight has a positive impact on profit, and the company is advised to consider total weight when setting pricing strategies for customers.
* Clustering Model: Utilizing the KMeans clustering algorithm, we performed customer segmentation based on order completion hour and ratings. The clustering analysis revealed five distinct customer segments, and customers with shorter order completion times tend to give higher ratings, indicating higher satisfaction levels.

Key Business Implications:

* The classification model's ability to predict political situations can support in making informed investment decisions for mining operations. Understanding the political stability of a region is crucial for assessing risks and potential returns.
* The regression model shows that company can use total weight i/o quantity as a key factor when setting prices for their services, potentially leading to increased profitability.
* Customer segmentation analysis can provide valuable insights into customer behavior and preferences. After discussing with the company, they doubt customers do not pay attention to the rating process after completing their order. Also, if customers do not give rating, the system assigns a default value that is why most of data spread in specific value.