1. First part of the research and development report:

Project Introduction and Background

This project aims to develop a demand forecasting model to help cement and aggregate manufacturers predict future market demand, so as to better plan production and sales plans.

In the field of construction and infrastructure construction, cement and aggregates are essential materials, so accurate forecasting of market demand is critical to the success of a business.

We selected two different materials for research and prediction, namely cement and aggregate. To develop this predictive model, we employed machine learning techniques and the Python programming language.

We will use historical sales data and other relevant data as input, train a model, and then use the model to predict future market demand.

The background of this project is the volatility of market demand and the problem of coordination between production planning and sales planning. If companies cannot accurately predict market demand, it will be difficult to plan production and sales plans, which may lead to problems of excess or insufficient supply. This has an impact on business productivity and profits.

Related tech-stack:

Numpy is an extension library for Python, dedicated to mathematical calculations. Numpy provides a convenient way to work with matrices and vectors. In this project, Numpy is used to process data sets and perform preprocessing such as data cleaning and feature engineering. Since Python is an interpreted language, efficient implementation of Numpy makes data processing more efficient.

Spark is a big data processing framework, which is a distributed computing framework that can process massive amounts of data. In this project, Spark is used for data preprocessing and model training. Because the data set is very large, the processing speed is very slow using traditional Python programs, but the distributed computing of Spark can greatly speed up data processing.

Bootstrap is a popular front-end framework that helps us build beautiful websites quickly. In this project, we use Vue and Bootstrap to build the front-end interface. Bootstrap provides many excellent CSS and JavaScript components, which can help us quickly build web pages, and make web pages have good visual effects and responsive layout.

Numpy provides efficient mathematical calculation and data processing capabilities, Spark provides large-scale data processing and distributed computing capabilities, and Bootstrap provides the ability to quickly build beautiful front-end interfaces. The advantages of these technologies make them ideal for use in this project

2. The second part of the research and development report:

Research Methods and Data Analysis

We used three different machine learning models to predict market demand, namely linear regression model, random forest regression model, and gradient boosted tree regression model. We implement these models using the Python programming language and the Spark distributed computing framework.

We obtained historical sales data from sales data for two different materials, which included timestamps, sales quantities, and other relevant data. We cleaned and prepared the data, including converting timestamps into time units such as years, months, and days, and used Python's Pandas library and Plotly library for data visualization and exploration.

We also feature engineered the data, converting timestamps and other relevant data into features that can be used to train the model. We used Spark's VectorAssembler function to combine these features into a feature vector.

Before training the model, we split the data into training and testing datasets. We use the training dataset to train the model and the test dataset to evaluate the predictive accuracy of the model.

we use three different machine learning models to predict future market demand and use an estimator to calculate the mean squared error for each model. We use Python's Plotly library and Flask framework to build web applications and present model prediction results

After the data preparation, we carried out data exploration and analysis, mainly focusing on the distribution and time trend of demand. By plotting histograms and boxplots, we can find that demand is not evenly distributed and there are some outliers. In terms of time trends, we used scatter plots to plot the relationship between demand and time, and found that there were cyclical and seasonal changes in demand. These results are very helpful for our subsequent modeling and prediction.

Next, we started modeling and forecasting. We selected three algorithms for prediction, namely linear regression, random forest regression and gradient boosted tree regression. we discover:

In the visualization chart of the prediction results, the Random Forest (Random Forest) model and the Gradient Boosting Tree (Gradient Boosting Tree)\

The prediction result of the model is closer to the actual demand value than the prediction result of the linear regression (Linear Regression) model,\

Select the model with the smallest mean square error, that is, the gradient boosting tree model, as the final inventory prediction model'

Finally, we built a simple web application using the Flask framework, allowing users to view our demand forecast results and data analysis reports. We also added some security measures, such as CSRF attack defense and SQL injection defense, to ensure the security and reliability of the application.

In general, this project provides us with a complete research and development process, from data preparation to data exploration and analysis, to modeling and prediction, and finally to the construction of web applications and the addition of security measures. In the process, we learned a lot about data science and machine learning, and also exercised our programming and teamwork skills

After completing data exploration, we need to perform data preprocessing. Data preprocessing is a necessary step that cleans the data, fills in missing values, handles outliers, etc. to make the data more accurate and useful.

In this project, our data set has been preliminarily cleaned, and there is no need to fill in missing values or handle outliers again. However, we need to process the timestamps in order to convert them into features that can be used for model training. We need to split the timestamp into features like year, month, day, day of week, and weekend, and use a vector assembler to combine these features into a single feature vector.

In this code, we use Spark's DataFrame API to split the timestamp into features like year, month, day, and day of the week. We also used the when() function to determine whether each date falls on a weekend, and used VectorAssembler() to combine these features into a single feature vector.

3. The third part of the research and development report:

After data preprocessing, we can start training the predictive model. We will use three different regression models to make predictions: Linear Regression, Random Forest Regression, and Gradient Boosted Trees Regression. We will use the Spark MLlib library to train and evaluate these models.

First, we will use a linear regression model to make predictions. We will use the LinearRegression() function to create a linear regression object and the fit() function to train the model. We will also use the transform() function to make predictions on the test data using the trained model and the RegressionEvaluator() function to calculate the mean squared error.

In this project, we used three different machine learning models (Linear Regression Model, Random Forest Model, Gradient Boosted Tree Model) to forecast concrete demand. After comparative evaluation, we finally chose the gradient boosting tree model as the final inventory forecasting model.

We evaluated the prediction performance of the model and used visualization charts to show the prediction results of the model. Through the comparison of the prediction results and the observation of the error distribution graph, we found that the prediction results of the random forest model and the gradient boosting tree model are closer to the actual demand value than the linear regression model. Ultimately, we choose the gradient boosted tree model as the final inventory forecasting model.

We generated feature vectors using feature engineering and fed them into the final gradient boosted tree model to obtain demand forecasts for 2023. The forecast results are saved in a CSV file, which can be viewed and analyzed by users.

At the same time, we also made predictions for three different models and compared them using the mean squared error as the evaluation metric. In the visualization chart of the prediction results, the prediction results of the Random Forest model and the Gradient Boosting Tree model are closer to the actual demand value than the prediction results of the Linear Regression model. Therefore, we chose the gradient boosted tree model with the smallest mean square error as the final inventory forecasting model.

We used feature engineering to generate feature data for 2023, and used the trained gradient boosting tree model to predict inventory demand. The prediction results are saved in a CSV file for later use.

We used a variety of data science techniques in our projects, including data preprocessing, feature engineering, model selection and evaluation, and more. Our inventory demand forecasting model can not only help companies better manage inventory, but also provide more accurate forecast results for companies, thereby helping companies make better decisions and improve their operating efficiency and competitiveness.

In this project, we used a variety of tools and technologies in Python, including NumPy, Pandas, PySpark, Scikit-Learn, and Flask, etc. These tools and techniques allow us to process large amounts of data and use a variety of machine learning models for prediction and analysis.

For data processing, we used Pandas and PySpark. Pandas is a Python data analysis library for processing and analyzing data. Its flexibility and ease of use enable us to quickly analyze and process data. PySpark is the Python API of Apache Spark, which can help us process large amounts of data and perform complex data analysis and processing.

In terms of machine learning, we used three different machine learning models, including linear regression, random forest regression, and gradient boosted tree regression. These models use different algorithms and techniques to make predictions, and each model has its own advantages and disadvantages. We use these models for forecasting and evaluate their forecasting performance, finally choosing the gradient boosted tree model as the final inventory forecasting model.

In terms of displaying prediction results, we used two Python data visualization libraries, Matplotlib and Plotly. Matplotlib is a Python plotting library for creating various types of static, interactive, and dynamic charts. Plotly is an open source drawing library that can create various types of interactive and dynamic charts. We use these two libraries to visualize forecast results, including error distribution plots, time series plots, etc.

In terms of implementing web applications, we use Flask, a Python web framework. Flask is a lightweight web framework for building web applications. We use Flask to create functions such as user system, personal homepage function and manage products to be analyzed, so that users can easily use our forecasting model for inventory forecasting and analysis.

Projects involve many different technologies and tools, including data processing, machine learning, data visualization, and web applications, among others. The integration of these technologies and tools enables us to complete a complete inventory forecasting and analysis system, providing malls with more accurate and efficient inventory management and decision support.

Conclusion:

In this project, we used various data science and machine learning techniques to predict the inventory consumption of building materials. We first used principal component analysis and BP neural network for inventory forecasting, and then used more complex machine learning algorithms, including linear regression, random forest regression, and gradient boosting tree regression, to predict building material inventory consumption.

Through the comparison of various models, we found that the prediction effect of the gradient boosting tree model is the best, so this model is selected as the final inventory forecasting model. We used Spark's big data processing capabilities to process larger datasets, enabling our model to handle more inventory data.

On the front end, we use Vue.js and Bootstrap to implement a beautiful and easy-to-use user interface, allowing users to easily register, log in, change passwords, etc., and manage the types of products that need to be analyzed. We also use Jupyter Notebook for algorithm integration, so that users can directly click on the Jupyter link on the website to view the solutions predicted by different algorithms. In the end, we used the Flask framework to realize the interactive data visualization of the website, allowing users to view the forecast results of building material inventory consumption through interactive charts.

the project covers a variety of data science and machine learning technologies, including data processing, feature engineering, model training and model evaluation, as well as front-end technology and website development technology. Through this project, we not only mastered various data science and machine learning techniques, but also learned how to apply these techniques to practical problems, and learned how to integrate algorithms into websites to provide users with interactive data visualization. These skills will be of great help to our future work and research