

# PROJECT PROPOSAL: PRODUCT DEMAND PREDICTION WITH MACHINE LEARNINGS

## TABLE OF CONTENTS:

- INTRODUCTION
- PROBLEM DEFINITION
- DESIGN THINKING
  - DATA COLLECTION
  - DATA PREPROCESSING
  - FEATURE ENGINEERING
  - MODEL SELECTION
  - MODEL TRAINING
  - EVALUATION
- CONCLUSION

## INTRODUCTION:

The problem is to create a machine learning model that forecasts product demand based on historical sales data and external factors. The goal is to help businesses optimize inventory management and production planning to efficiently meet customer needs.

## PROBLEM DEFINITION:

The problem of predicting demand for a new product based on its characteristics and description is critical for various industrial enterprises, wholesale and retail trade and, especially, for modern highly competitive sector of air transportation, as solving this problem will optimize production, management and logistics in order to maximize profits and minimize costs. We can do this using demand forecasting. The conventional demand sensing or demand forecasting methods assume the availability of sales data for a certain historic period. But in most scenarios that is not the case since historic data is not available for a new product. Demand forecasting is the estimation of a probable



future demand for a product or service. Demand planning serves as the starting point for many other activities, such as warehousing, shipping, price forecasting , financial planning, and, especially, supply planning that aims at fulfilling the demand and requires data on the anticipated needs of customers. Thus the problem solves the following issues:

- 1) Optimise production
- 2) Manage the logistics and retail to maximise profits and reduce costs
- 3) Inventory management : availability of products needed
- 4) Efficiently meet customer needs

## **DESIGN THINKING:**

### **1) DATA COLLECTION:**

There are 2 main types of data that needs to be collected. They are:

- a) Primary data
- b) Secondary data

There are different ways to collect these data. They are:

**Primary data:** Refers to the data that does not have any prior existence and collected directly from the respondents. It is considered very reliable in comparison to all other forms of data. However the reliability can sometimes be biased. Primary data would not be very dependable. Primary data collection should be done in an unbiased way and caution. This data helps the researchers in understanding the real situation of a problem. and presents the current scenario in front of the researchers and thus it is more effective in taking the business decisions.

## **METHODS OF PRIMARY DATA COLLECTION:**

### **1) OBSERVATION METHOD:**

Observation method is a method in which the population of interest is observed to find out relevant facts and figures. The observation method is further divided into 6 types.

### **1) NATURAL METHOD:**



Edit with WPS Office

It refers to the method in which the researcher observes the behaviour of people without any intervention.

## **2) CONTRIVED METHOD:**

It refers to the method in which the researcher takes the information from the people in an indirect way.

## **3) DIRECT METHOD:**

It refers to the method in which the researcher waits for a particular experiment or behaviour to occur.

## **4) INDIRECT METHOD:**

It refers to the method in which the researcher observes the behaviours that have occurred in the past.

## **5) STRUCTURED METHOD:**

It refers to the method in which the researcher knows what is to be observed.

## **6) UNSTRUCTURED METHOD:**

It refers to the method in which the researcher does not know what exactly he/she has to observe.

## **SECONDARY DATA:**

It refers to the data that is collected in the past, but can be utilized in the present scenario/research work. The collection of secondary data requires less time in comparison to primary data. There are 4 ways to collect it. They are:

### **1) COMPANY RECORDS:**

They provide the information in the form of balance sheets and sales records. This information is used to perform a trend analysis of data and forecast the overall growth of a company in future.



## **2) INTERNET:**

The Internet also provides lots of the data related to the research from different sources.

## **3) PRINT MEDIA:**

It gives you the information that is publicized. Print media includes newspapers, magazine, books, research papers, and journals. The data collected from print media is get an overview of the present market situation and expert's opinions on different topics.

## **4) CENSUS AND GOVERNMENT RECORDS:**

This data contains the personal information of respondents. It is used mostly used by government and big organizations.

## **DATA PREPROCESSING:**

Data preprocessing is an unavoidable step in every machine learning algorithm. It refers to cleaning and transforming and integrating the data.

### **WHY DATA PREPROCESSING?**

- Preprocessing ensures that the data is accurate and precise and seasonality is maintained
- It also ensures that the data is consistent and there are no breaks or shifts in demand level
- It makes sure that missing and noisy data are removed

### **STEPS INVOLVED IN DATA PREPROCESSING:**

#### **Getting the data set :**

To create a machine learning model we first need a dataset. The collection of data is known as dataset. Dataset that we use in ML are mostly as CSV(comma separated values) file. Sometimes also HTML or xlsx file. We are already been given with a data so we will proceed with that.

#### **Importing libraries:**

To perform the preprocessing in python we need some python libraries. These are required to perform particular function. These include NumPy,



Edit with WPS Office

Matplotlib, pyplot, Pandas.

### **Importing datasets:**

We already have the dataset, the next step is to import them into the currently working directory. We can use:

- 1)read csv()
- 2)uploading the file directly

### **Finding missing data:**

This is executed in two ways:

- 1)by deleting the particular row
- 2)by calculating the mean

We will use SCIKITLEARN library in our code (sklearn.preprocessing)

### **Encoding Categorical data:**

Since machine learning model completely works on mathematics and numbers, but if our dataset would have a categorical variable, then it may create trouble while building the model. So it is necessary to encode these categorical variables into numbers.

### **Splitting dataset into training and testing set:**

In machine learning data preprocessing, we divide our dataset into a training set and test set. This is one of the crucial steps of data preprocessing as by doing this, we can enhance the performance of our machine learning model.

### **FEATURE ENGINEERING:**

Feature engineering is the process of selecting and transforming relevant features from the raw data to improve the performance of ML models. In demand prediction for drugs on pharmacies, some of the most important features are:

- **Time-based features:** These features capture trends and patterns over time. Examples include day of the week, month, year, and holidays.
- **Store-based features:** These features capture pharmacy-specific characteristics. Examples include the location of the



Edit with WPS Office

pharmacy, the size of the pharmacy, and the customer demographics.

## Why Feature Engineering?

Feature engineering is a critical step in building predictive models for demand forecasting. Here are some key reasons why it matters:

1. **Enhanced Model Performance:**

Well-engineered features can capture underlying patterns and relationships in the data, leading to more accurate predictions

2. **Improved Interpretability:**

Feature engineering can make your models more interpretable. By creating meaningful features, you can gain insights into which factors are driving demand and how they impact your predictions.

3. **Handling Non-linearity:**

Real-world demand data is often non-linear, and feature engineering allows you to transform variables to better fit the assumptions of your chosen machine learning algorithm.

## STEPS INVOLVED IN FEATURE ENGINEERING:

1. **Feature Creation:** Feature creation is finding the most useful variables to be used in a predictive model. The new features are created by mixing existing features using addition, subtraction, and ration, and these new features have great flexibility.
2. **Transformations:** The transformation step of feature engineering involves adjusting the predictor variable to improve the accuracy and performance of the model. It ensures that the model is flexible to take input of the variety of data; it ensures that all the variables are on the same scale, making the model easier to understand
3. **Feature Extraction:** Feature extraction is an automated feature engineering process that generates new variables by extracting them from the raw data. The main aim of this step is to reduce the volume of data so that it can be easily used and managed for data modelling.
4. **Feature Selection:** While developing the machine learning model, only a few variables in the dataset are useful for building the model, and the rest features are either redundant or irrelevant. Hence it is very important to identify and select the most appropriate features



Edit with WPS Office

from the data and remove the irrelevant or less important features, which is done with the help of feature selection in machine learning. Feature selection is a way of selecting the subset of the most relevant features from the original features set by removing the redundant, irrelevant, or noisy features.

## **MODEL SELECTION:**

### **a) Regression model:**

- Regression models are also helpful in predicting future values from past ones.
- They can help determine underlying trends and deal with cases involving overstated prices.
- While quite rare in real-life business cases, we can see a linear correlation between the target feature that needs to be predicted and the rest of the available variables.
- Because of this, it is important to select the proper regression model based on the custom client's data.
- Regression models like Random Forest & XGBoost can also be used to forecast demand for the future

### **b) Random forest:**

- Random Forest is the more advanced approach that takes multiple decision trees and merges them together.
- By taking an average of all individual decision tree estimates, the random forest model results in more reliable forecasts.
- The model may be too slow for real-time predictions when analyzing a large number of trees.

Pros:

- 1)High accuracy
- 2)It can automatically identify seasonality and trends

Cons:



Edit with WPS Office

1)Needs more data points at least 30 (in this case) to perform this regression

## **MODEL TRAINING:**

Model training:

When training forecasting models, data scientists usually use historical data. By processing this data, algorithms provide ready-to-use trained models.

a) Validation:

- Model validation is a phase of machine learning that quantifies the ability of an ML or statistical model to produce predictions or outputs with enough fidelity to be used reliably to achieve business objectives.

- By using a cross-validation tuning method where the training dataset is split into several equal parts, training the forecasting models with different sets of hyper-parameters. The goal of this step is to figure out which model's parameters have the most accurate forecast

- This step requires the optimization of the forecasting model parameters to achieve high performance.

- By using a cross-validation tuning method where the training dataset is split into several equal parts, data scientists train forecasting models with different sets of hyper-parameters.

- The goal of this step is to figure out which model's parameters have the most accurate forecast.

B) Improvement:

- When researching the best business solutions, data scientists usually develop several machine learning models and then choose the ones that cover the project's requirements the best. -The improvement step involves the optimization of analytic results.

- For example, using model ensemble techniques, it's possible to reach a more accurate forecast. In that case, the accuracy is calculated by combining the results of multiple forecasting models.

## **EVALUATION:**



Edit with WPS Office



Evaluation in the case of machine learning, it is best practice. After we train our machine learning, it's important to understand how well our model has performed. Evaluation metrics are used for this

### **Regression Evaluation Metrics:**

Unlike classification, where we measure a model's performance by checking how correct it's predictions are, in regression we check it by measuring the difference in predicted and actual values, our objective is to minimize the metric score in order to improve our model. We will use the below example to understand more.

### **ROOT MEAN SQUARED ERROR:**

Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are metrics used to evaluate a Regression Model. These metrics tell us how accurate our predictions are and, what is the amount of deviation from the actual values.

Technically, RMSE is the Root of the Mean of the Square of Errors and MAE is the Mean of Absolute value of Errors. Here, errors are the differences between the predicted values (values predicted by our regression model) and the actual values of a variable.

$$RMSE = \sqrt{(\sum_{i=1}^n ||y(i) - y_1(i)||^2) / N}$$

### **MEAN ABSOLUTE ERROR:**

Mean Absolute Error (MAE) is a metric used to evaluate the accuracy of a model. It is calculated the sum of absolute errors divided by the sample size. The formula for MAE is:

$$MAE = \frac{1}{n} \sum |y_i - x_i|$$

where  $y_i$  is the observed value for the  $i$ th observation,  $x_i$  is the predicted value for the  $i$ th observation, and  $n$  is the total number of observations.

MAE is often used to assess the performance of regression models. It measures the average absolute difference between actual and predicted values. Unlike mean squared error (MSE), MAE calculates the error on the same scale as the data, making it easier to interpret.

To calculate MAE, you need to find all of your absolute errors, add them



up, and divide by the number of errors. Scikit-learn provides a function called `mean_absolute_error` that can be used to calculate MAE in Python.

## **CONCLUSION:**

In product demand prediction, machine learning algorithms can analyse historical sales patterns and predict future trends. The expectations of customers could not be met without strong supply chain that involves strategic planning. Proper demand prediction enables better planning and utilization of resources for business. Both qualitative and quantitative methods are available to help companies demand prediction better. Demand prediction helps businesses make informed decisions about a product. The inputs of renewal rates on each country helps to generate accurate results by count and rate based forecasts. Customers expect effective products and hassle free on time services. Since customer's expectations changing faster than ever, businesses need a method to predict demand accurately. As product forecasting plays a vital role in the business sector in every field, it helps in demand prediction. The solution for product demand prediction is a statistical and machine learning based solution that creates time series regarding each product and it's entitlement based on geographic locations

