Title: Predicting Future Sales - Forecasting Project

Author: Ayush Oturkar

Professor: Prof. Koulik Khamaru

Date: 15/01/2024

Introduction

I propose a project to develop a robust predictive model for accurately forecasting product sales across multiple shops, even under conditions of varying shop and product availability. This project will utilize a real-world dataset containing daily historical sales data, spanning January 2013 to October 2015, along with supplemental information about items, categories, and shops.

Data Source:

Link to Data Source -

https://www.kaggle.com/competitions/competitive-data-science-predict-future-sales

You are provided with daily historical sales data. The task is to forecast the total amount of products sold in every shop for the test set. Note that the list of shops and products slightly changes every month. Creating a robust model that can handle such situations is part of the challenge.

File descriptions

- **sales_train.csv** - the training set. Daily historical data from January 2013 to October 2015.
- **test.csv** - the test set. You need to forecast the sales for these shops and products for November 2015.
- **items.csv** - supplemental information about the items/products.
- **item_categories.csv** - supplemental information about the items categories.
- **shops.csv**- supplemental information about the shops.

Problem Statement

The challenge of this project lies in creating a model that can effectively handle the dynamic nature of shop and product availability. The dataset exhibits slight variations in the list of shops and products each month, necessitating a robust model that can adapt to these changes and maintain accurate predictions.

Objectives

The primary objectives of this project are as follows:

1. Develop a robust model for sales forecasting:
    - Explore various machine learning techniques, including time series analysis, regression, and potentially ensemble methods.
    - Consider model architectures that can inherently handle dynamicity in shop and product lists.
    - Prioritize prediction accuracy and model adaptability.
2. Evaluate model performance on test set:
    - Utilize appropriate evaluation metrics to assess the model's forecasting capabilities.
    - Analyze performance under different scenarios of shop and product changes.
3. Analyze model behavior and feature importance:
    - Understand how the model makes predictions and identifies key factors influencing sales.
    - Explore feature interactions and potential non-linear relationships.

Methodology

1. Data preprocessing and exploration:
    - Clean and preprocess the data to address missing values, outliers, and inconsistencies.
    - Conduct exploratory data analysis to uncover patterns, trends, and correlations within the dataset.
2. Model development:
    - Experiment with different machine learning algorithms and techniques.
    - Implement strategies to address the dynamic nature of shop and product availability, potentially incorporating feature engineering or model adaptation techniques.

    - Optimize model hyperparameters through cross-validation.
3. Model evaluation:
    - Evaluate the model on the test set using relevant metrics (e.g., RMSE, MAPE).
    - Conduct robustness analysis to assess the model's performance under different scenarios of shop and product changes.
4. Feature analysis and interpretation:
    - Utilize techniques such as feature importance scores or partial dependence plots to understand model behavior and feature contributions.

Expected Outcomes

- A well-performing predictive model capable of accurate sales forecasting, even with dynamic shop and product lists.
- Insights into key features and factors influencing sales patterns to an extent we can discern the predictability.

Tentative Timelines

Jan 16’ 2024 - Feb 15’ 2024:

Setting up the code base and infrastructure. Understanding the data in depth. Performing the Data Preprocessing and EDA over the data

Feb 15’ 2024 - April 1’ 2024:

Performing Feature Engineering and baseline model experiment. I will also be working on drafting the mid-term report. Will be sharing the mid term report by end of March 2024

April 1’2024 - April 15’ 2024:

Experimenting with different models, performing hyperparameter tuning. Focus on documenting the performance of models on the hold out set. Focus on model evaluation, interpretation of champion Model.

April 15’2024 - May 1’2024:

Completion of the final draft of the report.