|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Text  Description automatically generated with medium confidence** | | |  | | | **Master of Applied Computing** | | | |
|  | **ASSIGNMENT COVER SHEET** | | | | | | | | |
| Course Code: **ITS70604** | | | | | | | | |  |
| Course Name: **Data Analytics and Machine Learning** | | | | | | | | |  |
| **Student Name:**  Chen Tao PAN HUAIXIAN Hu Na | | | **Student Number:**  0353798 0351649 0353741 | | | **Group Number:**  4 | | |
| **Assignment No./Title:** | | | | Assignment-3 | | | | |
| **Course Tutor/Lecturer:** | | | | Dr. Riyaz Ahamed | | | | |

***Declaration***

*I/we\* certify that this assignment is entirely my/our\* own work, except where I/we\* have given fully documented references to the work of others, and that the material contained in this assignment has not previously been submitted for assessment in any other formal course of study.*

\* *Delete where applicable (depends on whether individual or group assignment).*

**Signature of Student**:

|  |  |
| --- | --- |
| **Grade:** |  |
| ***Evaluator's Comments:*** | |

Table of content

[Web based Sales Forecast System 3](#_Toc107152603)

[1. Problem Statement 3](#_Toc107152604)

[2. Dataset description and processing √ 6](#_Toc107152605)

[2.1. Dataset Description √ 6](#_Toc107152606)

[2.2. Data Processing //TODO 6](#_Toc107152607)

[3. Prototype design 7](#_Toc107152608)

[3.1. Prototype Workflow √ 7](#_Toc107152609)

[3.2. Component1 9](#_Toc107152610)

[3.3. Component2 9](#_Toc107152611)

[4. Architecture diagram and explanation 10](#_Toc107152612)

[4.1. Web Architecture //TODO 10](#_Toc107152613)

[4.2. Ensemble Model Architecture √ 11](#_Toc107152614)

[4.3. XGBoost 12](#_Toc107152615)

[4.4. Random Forest 12](#_Toc107152616)

[4.5. Linear Regression 12](#_Toc107152617)

[4.6. KNN 12](#_Toc107152618)

[5. Implementation and Model Evaluation - 13](#_Toc107152619)

[5.1. Feature Engineering 13](#_Toc107152620)

[5.2. Split Dataset 13](#_Toc107152621)

[5.3. Modeling - Tree Based 14](#_Toc107152622)

[5.4. Modeling – Linear Model 14](#_Toc107152623)

[5.5. Modeling – Clustering Model 14](#_Toc107152624)

[5.6. Ensemble Model 14](#_Toc107152625)

[6. Conclusion and Future Work //TODO 15](#_Toc107152626)

[6.1. Conclusion 15](#_Toc107152627)

[6.2. Future Work 15](#_Toc107152628)

[References 16](#_Toc107152629)

## Web based Sales Forecast System

## Problem Statement

Today's retail environment is more competitive than ever before. Buyers can compare prices in seconds, research your brand in minutes, and post comments that could lead to a domino effect that could benefit or hurt the retailer's brand and its sales (Akpinar, 2022)

Diagram1



Having exposure is no longer enough. In order to win customers and sell in the retail sector today, it is necessary to follow seven steps (Diagram 1) to carry out a competitive analysis of adjusted prices, personalise your marketing messages and your customers' shopping experience, and ensure that popular items are always available at the right price. There is only one way to do all this: Retailers can become more competitive by collecting and processing data and creating a circular workflow of price prediction -> pay response -> continuous collection, prediction and response

Our project uses a challenging time series dataset consisting of daily sales data provided by 1c company. The ultimate goal is to help 1ccompany solve the sales forecasting problem and thus help 1c company to better apply the model proposed by diagram1

**Purpose Statement & Objectives:**

Our ultimate aim in this research is to help 1c-company with their sales forecasting, so the context of our goal is to create a robust model that can handle time series data for price forecasting, while ensuring the application of the model and the collection of data and repetition of the training model. The problem will therefore focus on three areas: data processing, model building, and model evaluation.

Firstly our dataset has 5 datasets with different roles and the set of features is greater than 50, so the dataset faces the dimensional disaster problem.

When we have too many features, observations become harder to cluster and too many dimensions can cause each observation in the dataset to appear equidistant from all other observations. And because clustering uses a distance metric (e.g. Euclidean distance) to quantify the similarity between observations, this is a big problem. If the distances are all roughly equal, all observations look the same (as well as the same different) and no meaningful clusters can be formed (Yiu, 2021)

图表, 折线图

描述已自动生成

The second issue is how to choose the right model, which depends on the assessment method we choose.

In this application we want to built a framework that allows us to perform a forecast on sales using both algorithms and (ii) allow us to evaluate the best model numerically and graphically.

In order to do so, the following objectives should be achieved

1. The clear understanding of the procedure of a sales forecast and different evaluating tools
2. Dimensionality reduction of datasets, grouping by features
3. Create multiple models and evaluate them to select the best one
4. Create web application to help companies make better sales forecasts

**Prototype:**

Web

Model-ensemble

## Dataset description and processing √

### Dataset Description √

To estimate the following month's total sales for every product and shop. By participating in this contest, we will show to utilize and improve our data science abilities.

The data are given daily historical sales information. The assignment is to estimate the total number of goods sold in each store during the test set. Note that the list of stores and items varies significantly each month. Developing a model capable of handling such circumstances is part of the difficulty.

**Data fields description:**

\* ID - an Id that represents a (Shop, Item) tuple within the test set

\* shop\_id - unique identifier of a shop

\* item\_id - unique identifier of a product

\* item\_category\_id - unique identifier of item category

\* date\_block\_num - a consecutive month number, used for convenience. January 2013 is 0, February 2013 is 1,..., October 2015 is 33

\* date - date in format dd/mm/yyyy

\* item\_cnt\_day - number of products sold. You are predicting a monthly amount of this measure

\* item\_price - current price of an item

\* item\_name - name of item

\* shop\_name - name of shop

\* item\_category\_name - name of item category

### Data Processing //TODO

Fadfasdf

## Web App Prototype design

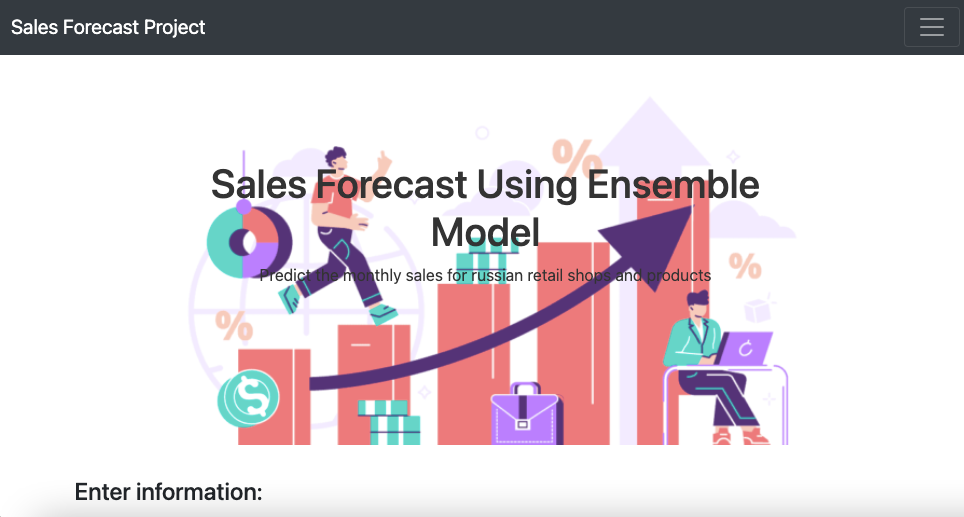
### Workflow √

Diagram

Description automatically generated

### Prototype √

**Main Page**



**Product Detail Input Page**

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application, chat or text message

Description automatically generated

**Prediction Output Page**

Diagram

Description automatically generated

### Component1

Balabla

### Component2

Balabla

## Architecture diagram and explanation

### Architecture

图示

描述已自动生成

1. Collect data

Our data will be divided into two parts, firstly the raw data provided by 1c-company, which is divided into 5 datasets, and secondly the data collected via the webpage, which is derived from the latest input from users

2. Data clean

In this section we check and process the data for null values and check for negative values, as it is unlikely that there are negative numbers in sales, and if there are, then it must be an abnormal input. Also some meaningless serial number features are removed.

3. Understand our data better in Exploratory Data Analysis, do necessary data wrangling

1. How sales behaves along the year
2. What category sells more
3. What shop sells more

At this stage we look at sales trends, best sellers and best shops over the year to help 1c-company make better sales strategies, such as increasing stock of best sellers, discounting products that are not selling well, etc.

4. Feature Engineering

1. Unitary item prices
2. Group based features.
3. How much each item's price changed from its (lowest/highest) historical price.
4. Rolling window based features (window = 3 months).
5. Lag based features.
6. Item sales count trend.

At this stage, we combined features and unified units to help with subsequent model building

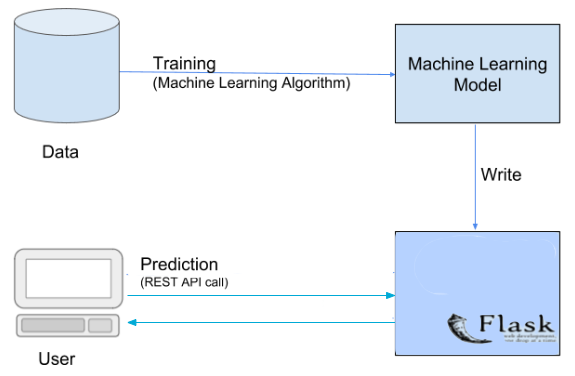
5. Set up Cross Validation to try out different feature engineering ideas

6. Use Ensemble methods to boost score

Ensemble learning is the process by which multiple models, such as classifiers or experts, are strategically generated and combined to solve a particular computational intelligence problem. Ensemble learning is primarily used to improve the (classification, prediction, function approximation, etc.)

he motivation for using ensemble models is to reduce the generalization error of the prediction. As long as the base models are diverse and independent, the prediction error of the model decreases when the ensemble approach is used. The approach seeks the wisdom of crowds in making a prediction. Even though the ensemble model has multiple base models within the model, it acts and performs as a single model. Most of the practical data mining solutions utilize ensemble modeling techniques.

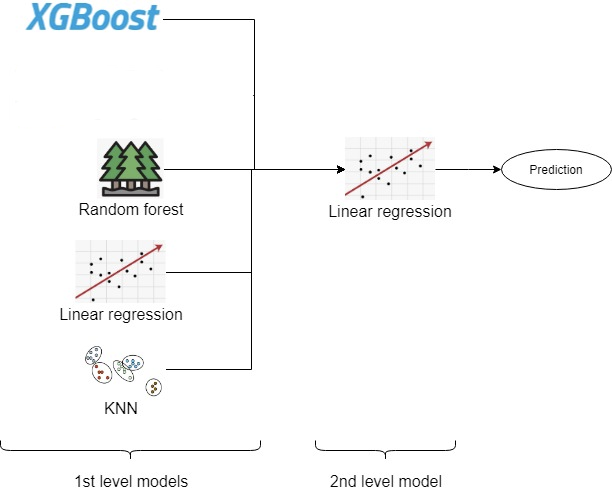
### Web Architecture //TODO



Use the api capabilities provided by html3/html5, such as the input component to collect information such as price and name and then pass the data to the back-end via REST API to complete the data collection task.

Visually we use the flex layout provided by html3, which automatically adapts to the size of the screen and ensures a good user experience.

### Ensemble Model Architecture √



Tree Based Model

XGBoost

RF

Linear Based Model

Linear Regression

Clustering Model

KNN

↓

Ensembled Model

### XGBoost

### Random Forest

blalaa

### Linear Regression

Blalaa

### KNN

Blalaa

## Implementation and Model Evaluation -

### Feature Engineering

### Split Dataset

### Modeling - Tree Based

### Modeling – Linear Model

### Modeling – Clustering Model

### Ensemble Model

## Conclusion and Future Work //TODO

### Conclusion

Result and Evaluation

Result:

Xxx RMSE

XXX application

### Future Work

Spark

Time Series Analysis

## **References**