**AIDI1002 – AI algorithms**

**Final Project**

**Sales Forecasting**



**NAME:** Dheeraj Reddy

**STUDENT NUMBER:** 100811270

**Prof:** Marcos Bittencourt

**Rational Statement:**

Many big stores need to understand the future sales to ensure that they have enough stock in their inventory. For a business to succeed it should never disappoint it’s loyal customers. If enough stock is not available for the customers they might find an alternative store, which will indirectly result in loss for the store. There is a possibility that the customer will start to prefer alternate options even for the future. So, it is important for a business to keep enough stocks, at the same time having too much inventory might also be a loss for the store. So, it is very important to estimate the inventory requirements for any occasion in an year.

**Problem Statement:**

So I would like to develop a sales forecast model which forecasts the sales of different stores. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality. The sales are also dependent on the week days, for example , people shopping on weekend will always be in a higher number than that of week days. Already thousands of individual managers are predicting sales based on their unique circumstances, the accuracy of results can be quite varied because of differed assessments. I would like to propose a sales forecast model to address all these issues and to increase the consistency and accuracy of forecasting, unlike that of manual forecast done by different people.

**Data requirements:**

To start developing the model, Rossmann Store Sales dataset will be used. It is a competition held 3 years ago and the dataset is available in Kaggle.

Rossmann operates over 3,000 drug stores in 7 European countries. Currently, Rossmann store managers are tasked with predicting their daily sales for up to six weeks in advance.

The features of the dataset are as follows:

* **Id** - an Id that represents a (Store, Date) duple within the test set
* **Store** - a unique Id for each store
* **Sales** - the turnover for any given day (this is what you are predicting)
* **Customers** - the number of customers on a given day
* **Open** - an indicator for whether the store was open: 0 = closed, 1 = open
* **StateHoliday** - indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. Note that all schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = None
* **SchoolHoliday** - indicates if the (Store, Date) was affected by the closure of public schools
* **StoreType** - differentiates between 4 different store models: a, b, c, d
* **Assortment** - describes an assortment level: a = basic, b = extra, c = extended
* **CompetitionDistance** - distance in meters to the nearest competitor store
* **CompetitionOpenSince**[Month/Year] - gives the approximate year and month of the time the nearest competitor was opened
* **Promo** - indicates whether a store is running a promo on that day
* **Promo2** - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating
* **Promo2Since[Year/Week]** - describes the year and calendar week when the store started participating in Promo2
* **PromoInterval** - describes the consecutive intervals Promo2 is started, naming the months the promotion is started anew. E.g. "Feb,May,Aug,Nov" means each round starts in February, May, August, November of any given year for that store

Link for dataset : <https://www.kaggle.com/c/favorita-grocery-sales-forecasting/data>

**Assumptions and Limitations:**

1. Data exploratory and Principle Component analysis will be done to identify keay features effecting the sales of the store.
2. All the null values will either be deleted or imputation assumption will be done if the null rows are in large number
3. This dataset is based on the European drug market and may not be suitable for North American market. However some learnings can be taken from this to train on any North American dataset.
4. Since a regression model is needed to be built, the output will be a float value, however it will be rounded off to a nearest integer to appropriately predict the number of sales

**Test Metrics:**

Since this is a Regression model, the root mean square error will be used as a primary metric.

However to check where do I stand in the leaderboard on Kaggle root-mean-square-percentage error will be used.

Target metric RMSPE < 10% on the test set.