Predict Future Sales - Data Preparation and Exploratory Data Analysis

This is an EDA project using python based on Kaggle competition: Predict Future Sales

In this competition, time-series dataset consisting of daily sales data, is provided by one of the largest Russian software firms - 1C Company. We are provided with daily sales data for each store-item combination, but our task is to predict total sales for every product and store in the next upcoming month.

FILES DESCRIPTION

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.
Sample_submission.csv	a sample submission file in the correct format
items.csv	supplemental information about the items/products.
items_categories.csv	supplemental information about the items categories.
shops.csv	supplemental information about the shops.

DATA 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
item_cnt_day	number of products sold. You are predicting a monthly amount of this measure
item_price	current price of an item
date_block_num	a consecutive month number, used for convenience. January 2013 is 0, February 2013 is 1,..., October 2015 is 33
item_name	name of item

shop_name	name of shop
item_category_name	name of item category
date	date in format dd/mm/yyyy

Data Preparation and Exploratory Analysis

Following steps were performed for data preparation and analysis to get the best predictions.

STATISTICALLY EXPLORING DATA

Checked:

- Data dimensions, rows and columns
- Columns names
- Data Type and data Info
- Unique values per column (to decide whether to make column Categorical?)

CLEANING DATA

- Look for any missing Data
- Identify and convert Categorical columns/values to Numerical representation using Dummy Variables if suitable for modelling
- Identify and convert Numerical columns/values to Categorical representation
- Check for distinct values in Categorical columns

STATISTICAL OVERVIEW OF DATA

- Check head, tail of data to see complete required data loaded
- Describe data, columns
- Identify numerical columns and look for insights like mean, median, mode, etc.
- Understand the relationship of columns and how they are affecting each other.
- Apply feature engineering by creating new columns which may help in better understanding of data and predictions (grouped the date into weekday, month and year)

RESULTS

There is a clear weekly cycle, with more sales on Friday, Saturday and Sunday with Saturday being the favorite shopping day. When I grouped sales by month, we can clearly see that: December and January had the highest item_cnt_day. December also had the highest item_sale (most likely as a result of christmas) November had the lowest item_cnt_day but July had the lowest item_sale.

Also, there are a few peaks over the years, they are:

- End of November 2013, lots of revenues and item_cnt_day.
- End of December 2013 had relatively lower item_sale compared to Nov 2013 and End of Dec 2014 where there was a high number of item_cnt_day.
- There are also peaks around the end of May 2014 and 2015. We also see declines in item_cnt_day and item_sale from 2013 to 2015.

By item, the most popular were:

- By item_cnt_day was: **item_id 20949,** with 187642 units sold, generated $929k
- By item_sale was: **item_id 6675**, worth $219M in revenue, with 10289 units sold
- The worst are: item 1590, 11871, 18062, 13474, 13477. Shops lost money on them.

By shop: shop_id 31 had the highest item_count day(310777 items sold), and also highest item sale($235M). **Shop_id 36** had the lowest item_cnt_day(330), with a revenue of ($377k)

Item_category: The most popular by item_cnt_day is **item_category 40**, with 634171 units sold, generating $170M. While by item_sale is **item_category 19**, with 254887 units sold, generating $412M . The worst are **item_category 51**, with only one unit sold ($129), and **item_category 50**, with 3 units sold ($24).

CONCLUSION

- Most of the shops have a similar selling rate, but 3 of them have a much higher rate, this indicative of the shop size or location.
- Better awareness on item _category 51 and 50 should be done because sales of the products in this category is very low which may be as a result of poor advertising. Also the products here could be paired with products in categories that are doing well (can be in form of promotion) to increase sales.
- These products that aren't bringing in revenue can be scrapped because of the cost incurred in keeping them.
- Transferring of cost spent on idle products to active products can help increase revenue

- Some shops are not in operation as they sell very few to zero items so they should be closed down to avoid unnecessary expenses.
- Generally revenue, sales, item_cnt_day have been worse in 2015. There are more shops that aren't making sales and a reduction in the item_cnt_day , which shows a reduction in supply of products. Distribution of revenue generation is spread among fewer item categories.
- Also noticed that the shops that have low revenue are not selling the items from categories that have high selling rate.