Case Assignment with EY - Data & Analytics

Assignment 1: Data Science

Prepared by Tiancheng Qu

Recommendations

- Increase stock for Games root category(includes both PC and Console subcategories) items at all shops.
- Increase stock for Tickets root category(includes both Ticket(digits) and Utilities-Tickets) items at all shops.
- Increase stock for Gift root category items at all shops.
- *‘Khimki TC "Mega"’, Moscow TC "Perlovskiy“, Krasnoyarsk TC "Vzletka Plaza“, Ufa TC "Central“, Mytishchi TRK "XL-3“, St. Petersburg TK "Sennaya“, Yakutsk Ordzhonikidze, 56, RostovNaDonu TRC "Megacenter Horizon" Island and Tyumen SC "Goodwin“* shops predicted to have significant decrease in revenue. It is worth looking into the store management and or other store performance matrix.
- Game Console promotions may be needed for *Shop Online Emergencies, Moscow shopping center "Semyonov“, and Moscow shopping center "MEGA Teply Stan" II of.* Since they are predicted to suffer major revenue lost.

Approach-Process

- Python and its various libraries were used as the scripting language.
- Jupyter Notebook was used to preform the coding, visualization and analytics.
- Use sales record as the base to create an information rich dataframe that joins the information from item, item category and shops that are active.
- Generalized various of similar item categories into one root category
- Check and correct any ‘incorrect’ information such as unexpected item price.
- Group the sales data by shop name and item’s root category
- Utilizing time series analysis and prediction models (SARIMA, ARIMA) to make subsequent month (2016-01-01) shop root category sales predictions.
- The model with the better test set performance will be the one that is making the sales prediction in each case.
- Identify the shop-root category groups where predicted to have lower sales then pervious years at same month
- Create recommendations based on the root category characteristics

Approach-Assumptions

- Assume inactive is represented as 'active_flag' ==‘X’
- Assume generalized root categories are sufficient for making ‘detailed’ prediction for total sales for ‘every product’ and store.
- Assume negative value of item_cnt_day means return of the product and full refund was given.
- Assume empty entry or item_cnt_day indicate no sale or return of such product.
- Assume shop with the root category that has no sales record for 24 months before the end of the sales record are not active and no prediction will be generated.
- Assume extreme high ‘item_price’ during October 2013 were data entry error.

Key Actions

1. Excluded any row that was marked as inactive.
2. Translated ‘Игры’ to ‘Games’ in ‘item_category_name’
3. Upon inspected the ‘item_category_name’ column, ‘item_root_categories’ was created and the values were manually added to the file based on the similarity between each category name.
4. Join the item dataframe with item category dataframe.
5. Join the sales dataframe with shops dataframe.
6. Join the results from Action 3 and 4 to create the comprehensive dataframe, referred as ‘sales_df_working’.
7. Multiplied the ‘item_price’ with ‘item_cnt_day’ to create the column ‘revenue’ in the ‘sales_df_working’.
8. Calculated and visualized each shop’s monthly revenue, using aggregation.
9. Large spike of revenue for almost all shop during October 2013 was observed. Upon close inspection, some of the item_price were abnormally high even compare with its price history within the sales record. Is it likely due to human entry error.
10. Further research into the true possible item price online confirmed the assumption, a threshold of ₽6000 was determined and any item price that is higher will be divided by 100.
11. The sales record dataframe with adjustments was visualized again to verify that the effect of data entry error has been reduced.
12. Two types of times series model were selected (SARIMAX and ARIMA)
13. shop with the root category that has no sales record for 24 months before the end of the sales record are considered not active and no predictions were generated.
14. The model with the better test set performance will be the one that is making the sales prediction in each case.
15. All models’ best parameters and predictions were saved in dataframe for reproduction.

Model Evaluation and Results

- There were 704 total possible store name and root category combinations.
- The models were able to make 488 predictions.
- SARIMAX was up to 2x better at predicting the sales winning 331 times, vs. ARIMA at 157 times.
- Based on the predicted store and root category pair, we are expecting to earn ₽15M more revenue compared with 2015 January.
- 9 out of 50 shops are expected to make less revenue compared with 2015
- *Khimki TC "Mega“* is predicted to lost the most amount of revenue (- ₽2.47M). Revenue lost in every root categories. This could be the problem with the model or the store, further investigation required.
- *Moscow TRC "Atrium“* is predicted to generate the most amount of revenue (+₽2.51M)
- The predicted revenue gain leaders are: Games, Tickets, Gifts and Delivery
- The predicted revenue loss leaders are: Game consoles, Payment Cards, Movies and Programs

Case Assignment with EY - Data & Analytics

Thank You