Capstone Report

Problem Statement:

Predicting next month sales by item, by shop for a Russian gaming retailer using 34 months of sale history.

Benefits

Main audiences: Leadership of the gaming retailer.

They would leverage the results to decide the level of inventory investment and store volume allocation strategy to maximum in-stock level, hence maximizing sales. Leadership may use the result to determine needed marketing activities to further increase sales potentials.

Secondary audiences include:

- **Toy product developers**: Study may reflect meaningful category trend in which developer can prioritize resources on innovation in those categories
- **Competitors**: leverage the result to develop market/product strategy to tap into market potential
- **Data Scientists/Researchers**: If the model generates reasonable prediction, it can be scaled up to include more company's data and to be used to predict total market share for the toy industry

Dataset Description

Obtained from Kaggle, data are in form of 4 dataset

- Categories: include category names and respective id (size: (84, 2))
- Items: include product name, and respective id, and category id (size: (22710 x 3))
- Shops: include shop name and respective id (size: (60,2))
- Sales: main training dataset, have attributes date, date block number (equivalent to month), shop id, item id, item price, and item count daily

Cleaning and Wrangling

- Check and do not find missing values
- One price outlier identified with a price tag of $300K+, drop from the dataset
- One negative price item was found at one location, replace that with price from the same product but sold in a different shop
- Convert date from string to DateTime recognizable format
- Merge with categories, items, and shops dataset to get descriptive information for exploratory analysis
- Save final data frame into a new csv "train_df.csv" for ease of access

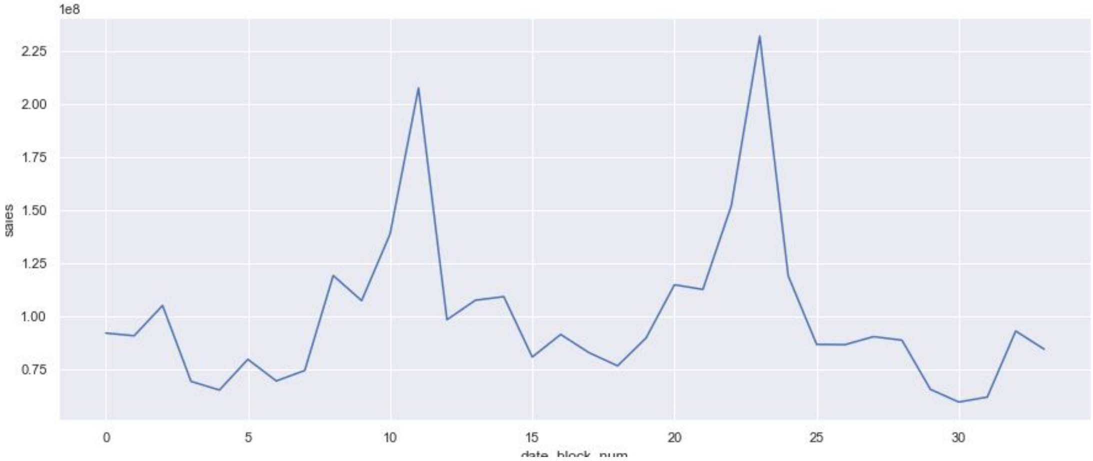
Exploratory Analysis

The goal is to discover features that may have predictive power to target variables (i.e. sales) through visualizing their historical pattern.

Areas of focus:

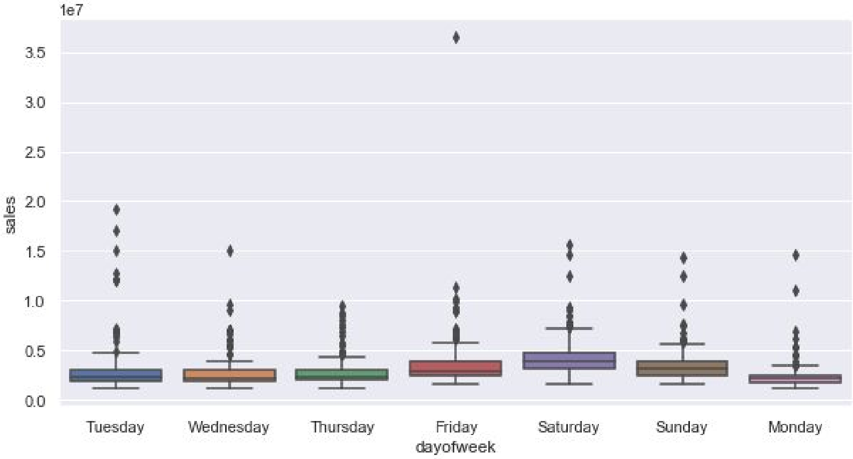
- Is there a seasonal pattern on sales in general?
- Would the day of the week (e.g. Monday vs Friday) be consistently different?
- Would sales be influenced by public holidays such as Christmas?
- Would there be certain products that dominate sales volumes?
- Would there be shops that have bigger shares of total sales?
- Would shops that behave similarly that can be grouped together for analysis?

Seasonality:



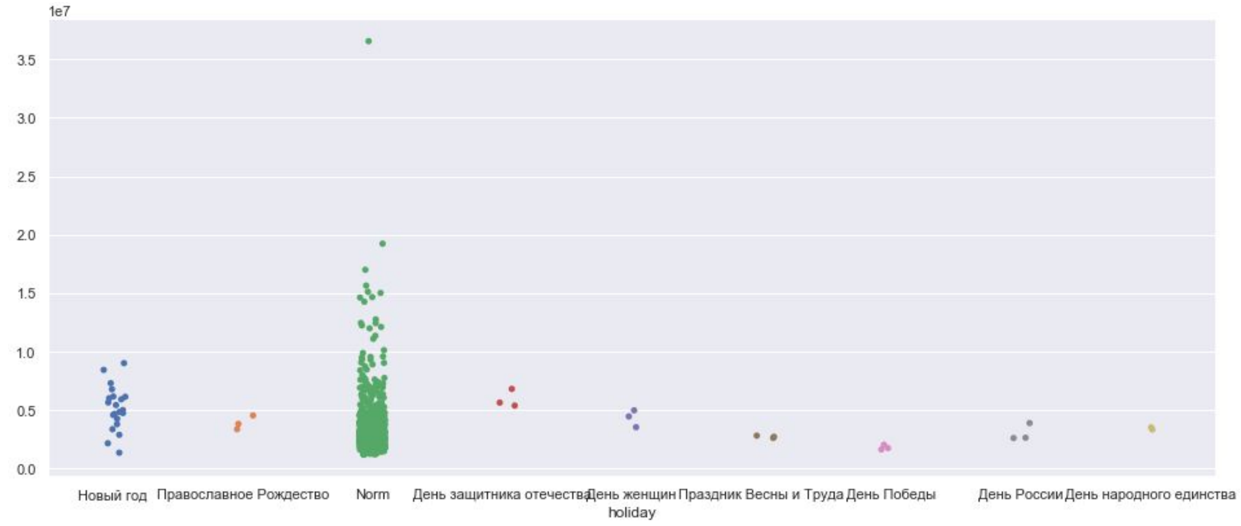
- Appears to have recurring sales pattern, with sales spike happens at every 12 months

Day Of Week



- Saturday and Sunday appear to have higher sales

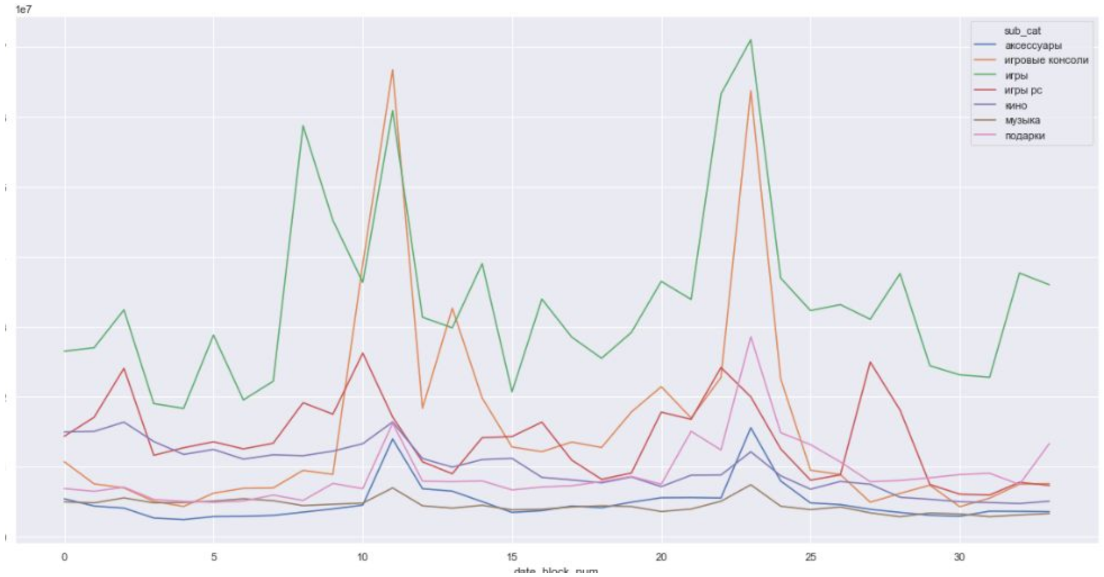
Holiday



- Not enough data points to justify any pattern

Product Category

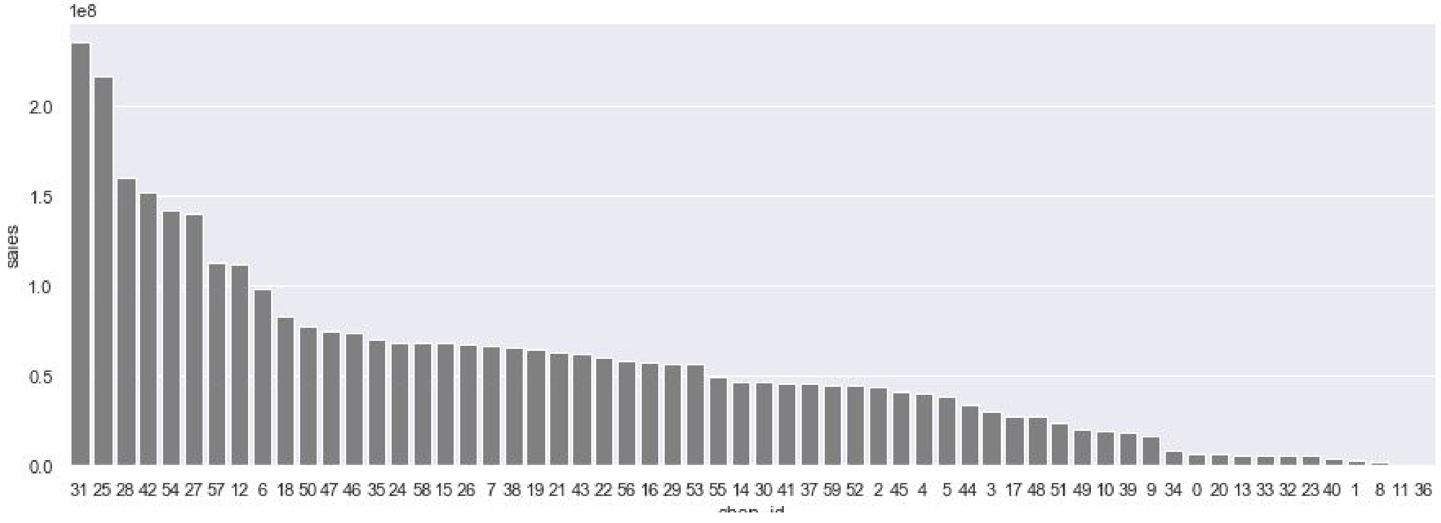
- Top 7 categories contribute to 90% of total sales
- Created a higher category based on product category name affix. Categories trend to be different over time



- games category (игры) seem to have more sales spike throughout the year. This is probably the major sales driver for the company and would tie to the game launch schedule.
- games pc (игры pc) has some fluctuations but the range is much smaller.
- game consoles ('игровые консоли') sub cat only spike during December and is pretty stable for the rest of the year. That's a good insight as it makes forecasting for this category easier outside December. The same insight applies to accessories (аксессуары) and gifts (подарки) category
- movie (кино) appear to be a declining sub cat and will need special attention
- aside from small spikes in December, music (музыка) sub cat is mostly flat.

Shop

- There is a total of 60 shops, but sales general skewed toward top 9-10 stores



Summary

From the above analysis, we identified a seasonal sales pattern, hence month of the year play a role in sales volume. Attributes such as product category and shop could play a big role in predicting sales since the majority of sales happen at top 7 category and top 9-10 stores

Next steps will be to apply statistical tests on time series patterns and significant value for the variables under regression.

Statistical Test

For access time series nature of the data, we will apply the **Dickey-Fuller** Test

H0: a unit root is present in an autoregressive model

H1: data are stationarity/trend-stationarity

p-value is **0.0088%.** We will reject H0 under 95% confidence level and favor the alternative that data are stationary.

To access whether the attributes in the dataset hold predictive power to the final target (item_cnt_day). We run OLS on all the attributes and access p-values for those

Attributes include item category, shop, month of the year and item prices. Results as follow:

Attributes	**p-value**
item_price	0.041
shop_id	avg 0.013
item_category_id	avg 0.026
month	avg 0.127

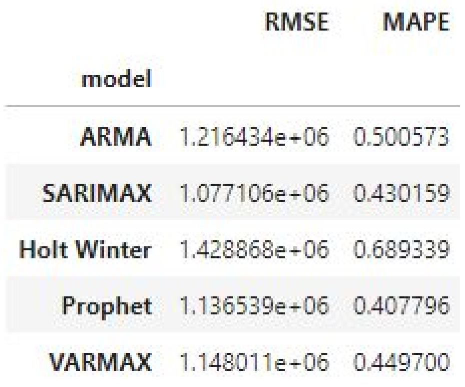
While month attributes average to heave high p-values, 9 out of the 12 months show almost 0 p-values. We will keep months as feature candidates. Also, we didn’t test item_id due its long list of values, we will assume that it has predictive power and item_price and item_category_id show significant importance with low p-values.

The final training dataset will include item_price, item_id, item_category_id, shop_id, and month as features

Machine Learning

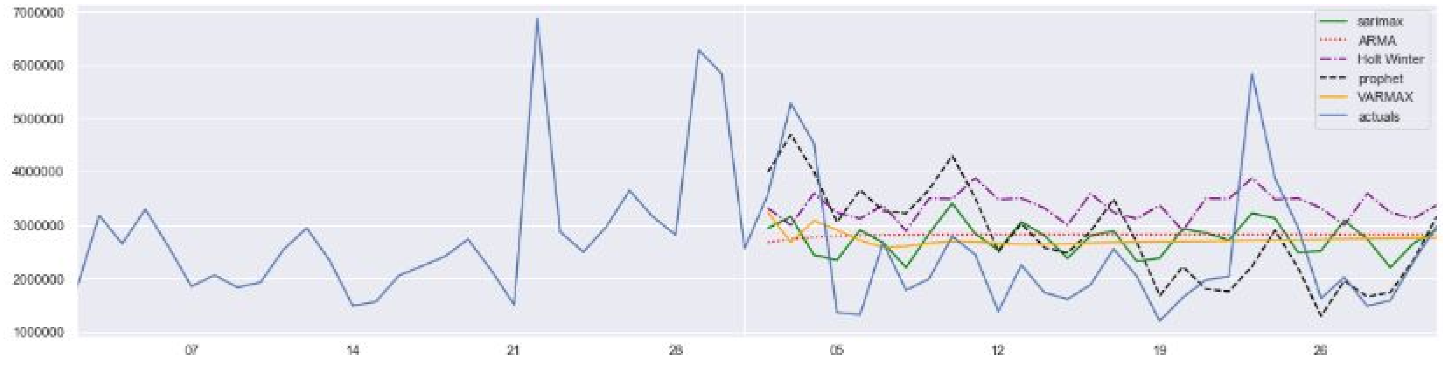
Time Series: models tested include ARMA, ARIMA, SARIMA, Prophet, Exponential Smoothing, VARMAX

Training Results:



	RMSE	MAPE
model		
ARMA	1.216434e+06	0.500573
SARIMAX	1.077106e+06	0.430159
Holt Winter	1.428868e+06	0.689339
Prophet	1.136539e+06	0.407796
VARMAX	1.148011e+06	0.449700

Graphically:



Observations: all the time series models do not have good performance. From the graphical results, we also see that there are days that the projections were way off (MAPE > 40% off). We would make an initial conclusion that time series models are not the best model on predicting sales at the daily level. The errors will likely get bigger as we try to drill down to store/product level.

Classification/Regression: modeled tested include Xgboost, Random Forest, Ridge Regression

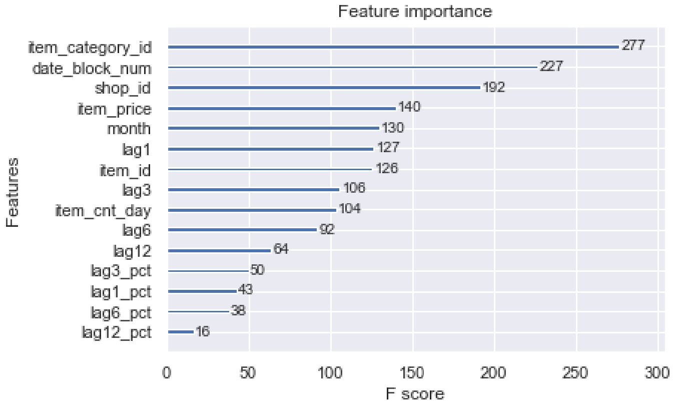
Custom Features: item_cnt_month lag(1, 3, 6, 12) and percentage change of item_cnt_month comparing its (1, 3, 6, 12) lags.

Training results:

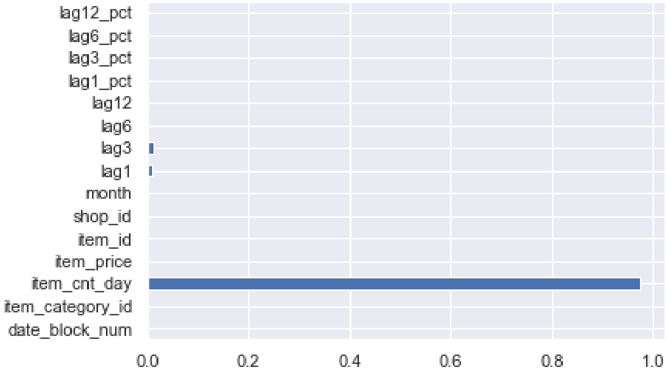
Models	**RMSE**
Xgboost	3.91
Random forest	3.85
Ridge	4.03

Observations: Classification/regression work much better for the predictions, all three models perform at a similar level. Note that the feature importances between xgboost and random forest are quite different. With the close results among the three models, we will submit them all to final hold out test validation by Kaggle

Xgboost



Random Forest



Kaggle Submission Results

Models	Naive	SARIMA	Xgboost	Random Forest	Ridge
RMSE	3.77	12.40	2.16	1.40	3.42

Naive model uses the same month the year before as submission (i.e. Oct 2014 to predict Oct 2015). Confirmed that time Series models such as SARIMA do not serve well in this situation.

Recommendation/Next Steps

Random Forest has the best performance on RMSE accuracy among all. From the execution standpoint, xgboost appears to run faster.

Next steps will be to include more relevant features at the shop level trend and product category trend to improve model performance. Yet, the trade-off of getting more features may lead to overfitting and more costly to run. Clients will need to decide what RMSE level they would feel comfortable enough on making business decisions.

It would be helpful if Clients can provide information on company internal marketing efforts (e.g. markdown/promotional activities) which help generate other useful features that are absent in the current dataset.