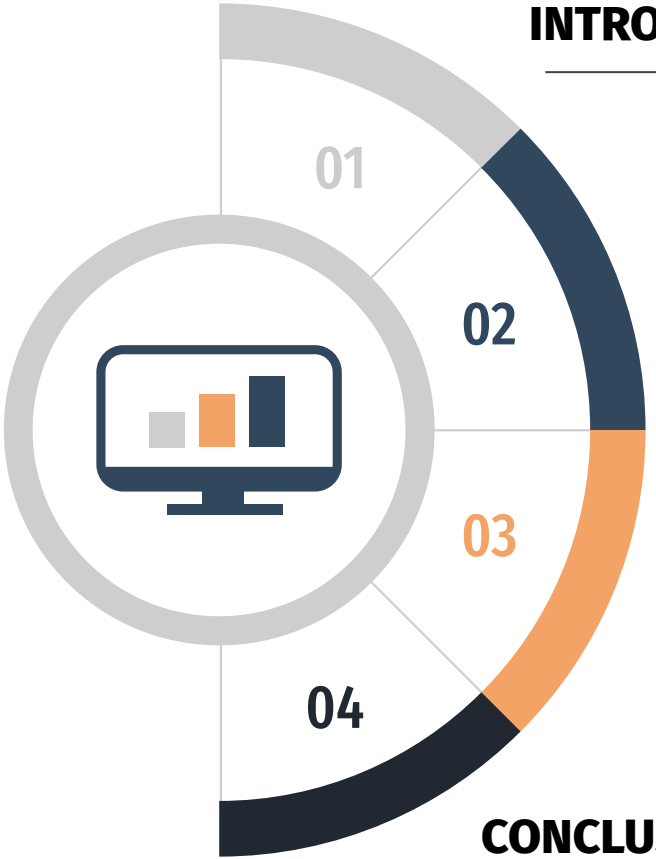
DSI30 CAPSTONE PROJECT

Predicting Sales: Time Series Analysis and Forecasting

8 Sep 2022

AGENDA



INTRODUCTION

EXPLORATORY DATA ANALYSIS

MODELLING

CONCLUSION

INTRODUCTION: The Company

Moscow, Russia

2016 Annual Revenue:
37 billion rubles (USD $650 million)

Expertise:

- **Development**
- **Manufacturing**
- **Licensing**
- **Support**
- **Distribution**

Products:

- **Computer software**
- **Video games**

PROBLEM STATEMENT

METRICS

Root Mean Squared Error

THE TASK

Time-Series Forecasting,
Regression Model

APPROACH

Bottoms-up Sales Forecasting
in Units

AUDIENCE

Decision-makers and Executives,
Planning Department

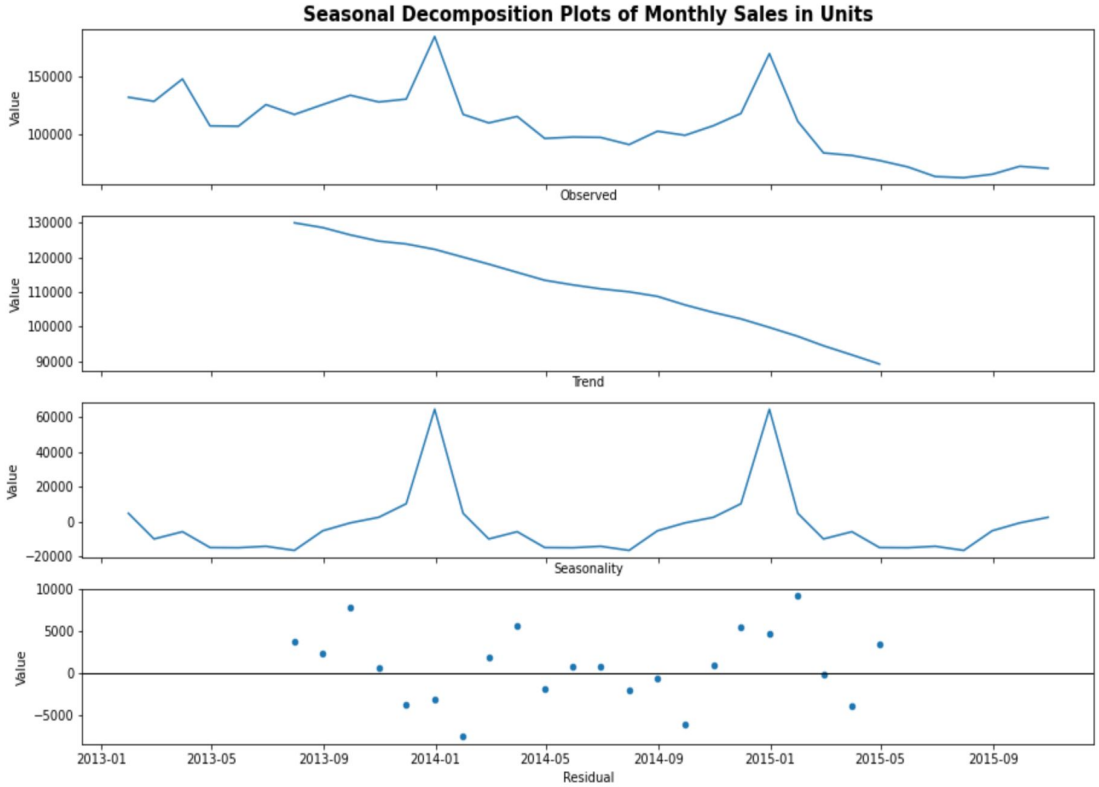
BUSINESS IMPACT

Maximize revenue
opportunities, minimize risk

BENEFITS

Reasonable accuracy, shorter
processing time

EDA: Unit Sales Trend and Seasonality



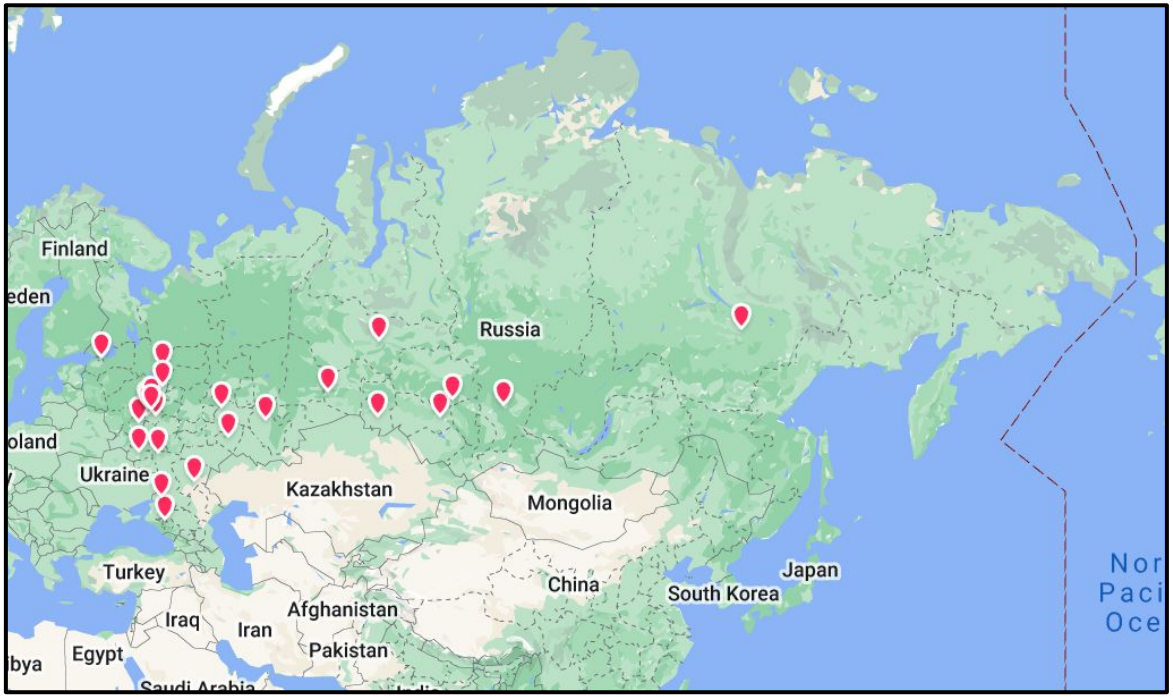
Yearly peaks in January

- Corresponds with Russian festive season

Downward trend in unit sales

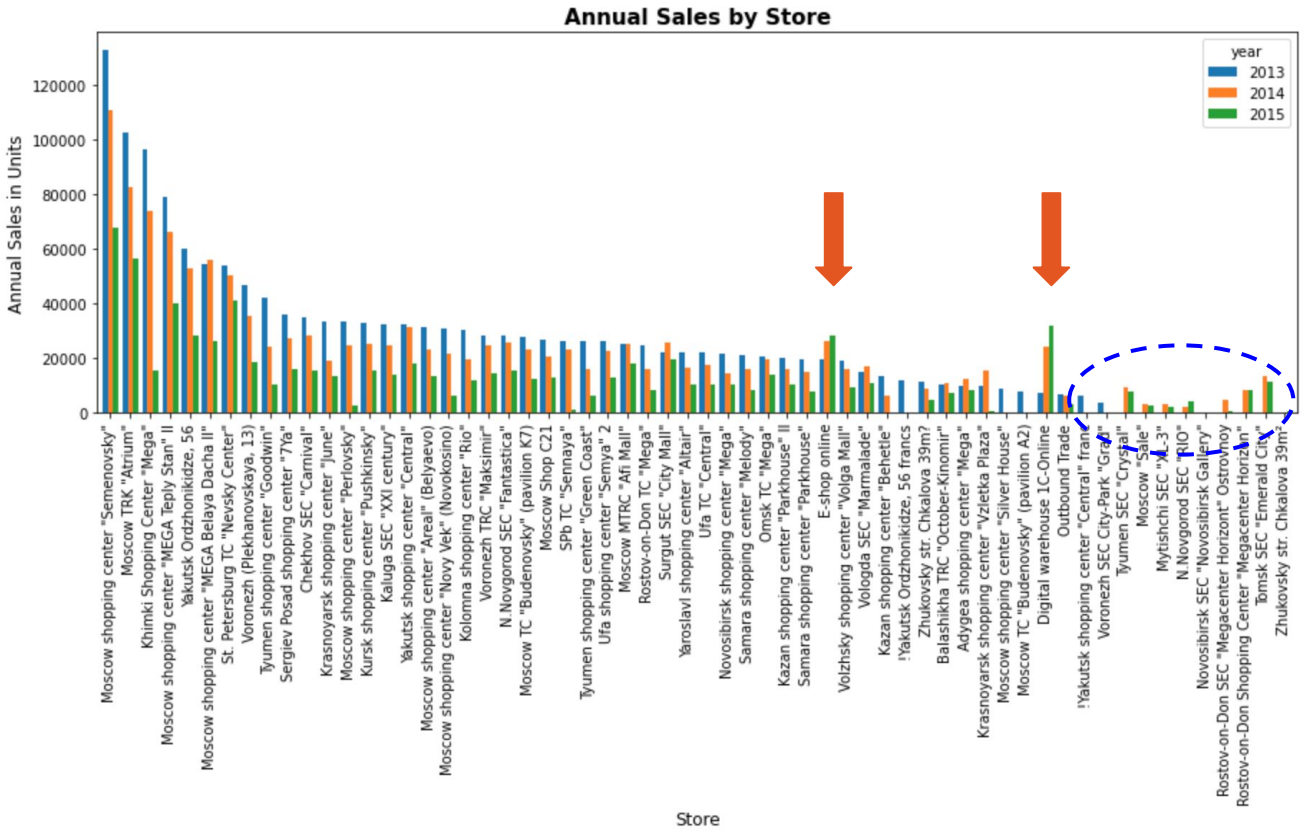
- However, sales revenue is increasing YoY

EDA: Geographical Distribution of Stores



- Customer demographics vary by geographical location.
- Demographics influence sales behaviour
- The more diverse the locations of the sales outlets, the more variations we need to consider during forecasting.

EDA: Sales Trend by Store/Channel



3 CHANNELS

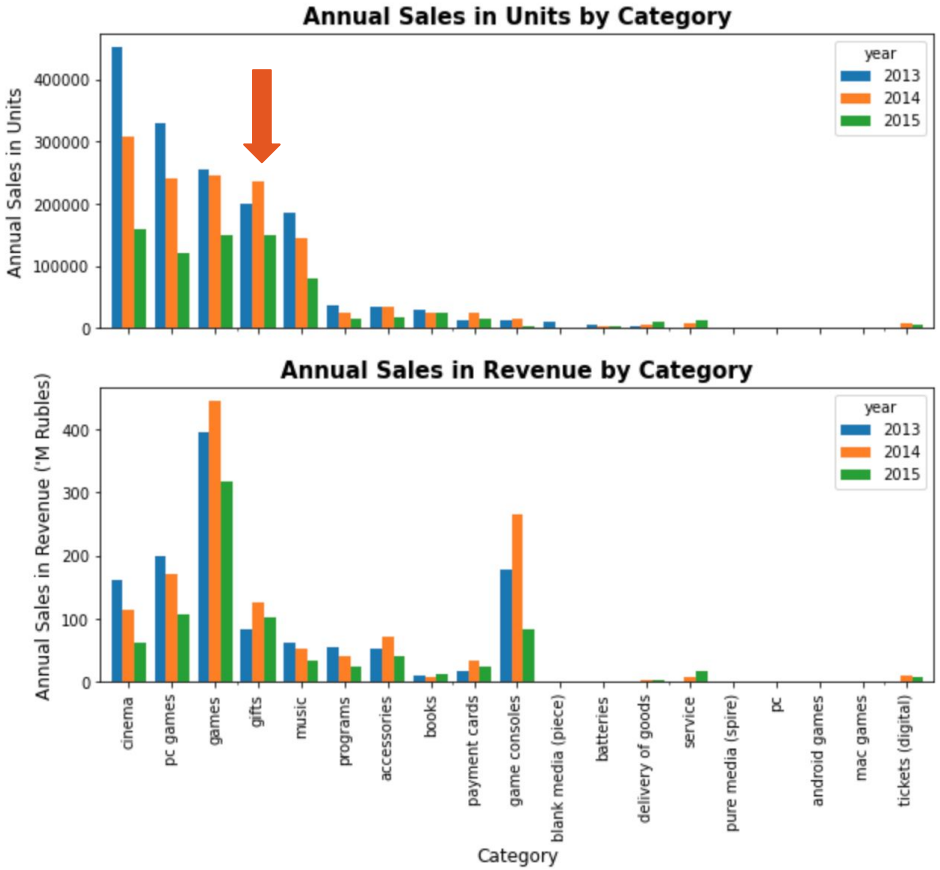
28 CITIES

60 OUTLETS

COMPLEXITIES:

- Different growth trends across shops
- Some shops not active in all years

EDA: Sales Trend by Item Category



18 Categories

59 Sub-categories

Item Counts:

2013: 14,967

2014: 14,105

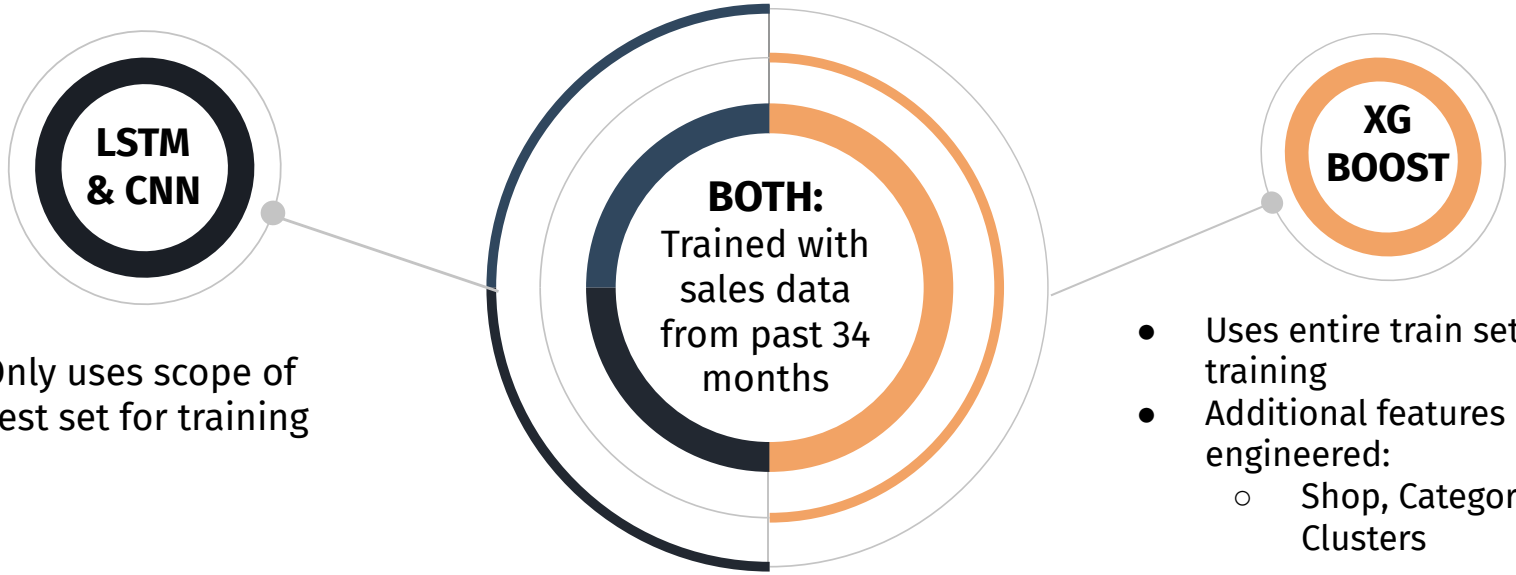
COMPLEXITIES:

- Different behaviour for different categories of products
- Assortment is different by store and period

MODELLING: Overview of Workflow

1. Data cleaning, feature-engineering
2. Train-test split
3. **Modelling**
   - LSTM
   - CNN
   - XGBoost Regressor
4. **Evaluation**
   Root Mean Squared Error

MODELLING: Approach



- Only uses scope of Test set for training

- Uses entire train set for training
- Additional features engineered:
  - Shop, Category Clusters

Dataset	Stores	Categories	Sub-Categories	Item Counts	Rows	Period
TRAIN	59	18	68	21,803	424,097	Jan'13-Oct'15
TEST	42	16	51	5,100	214,200	Nov'15

MODELLING: Evaluation

Profile of Test set:

Shop-Item Combination	%
No match in Train set	86.1
Only data from 1 year available	11.6
Data available in both 2013-14	2.3

Baseline score: 1.21756

- Based on a top-down estimate
- Average of Nov'13 and Nov'14 total sales, equally distributed across all rows in the test set

MODELLING: Results

Model	RMSE	Aggregate Sales
Baseline	1.21756	56,991*
LSTM	1.05313	22,154
CNN	1.04483	29,157
XGBoost Regressor	1.03877	27,642

**average of Nov'13 and Nov'14 sales*

- Model with the **best results** - ie: lowest RMSE - is the XGBoost Regressor
- Baseline model: a single **float** distributed equally across all shop-item combination
- Trained models: evaluated based on results that were **rounded to full integers**

CONCLUSION: Solving a complex problem...

Product Characteristics

Limited availability of historical data

Seasonality

Trend

Processing Time

Location

Proposed Future Works

Fine-Tune Features

Eg: Other approaches that can be used to create clusters

Optimize parameters

Eg: Threshold to round predictions to full integers

Explore other models

Eg: Hierarchical Time Series

Thank You!

Any questions?

Inspiration for the journey ahead:

“You may not like video games. But what I learned from them is this: no enemies in front of you means you are going in the wrong way.”