Future Sales Prediction

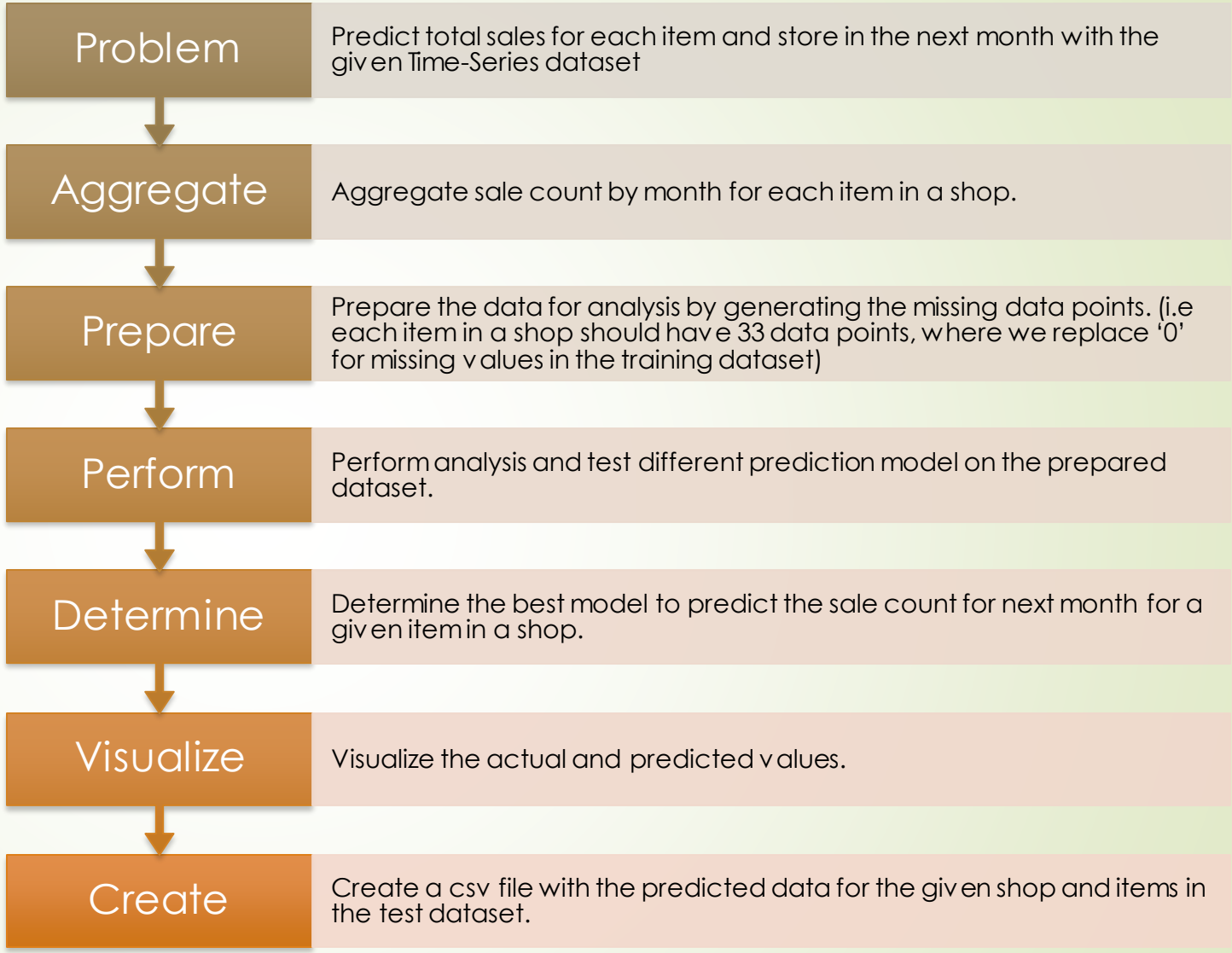
Presented by

Charannag Devarapalli
Deepti Gupta
Madhu shri Rajagopalan
Pallava ArAsu Pari
Yash Anilbhai Shah

INTRODUCTION

- Machine Learning techniques are used by all businesses in future prediction of sales.
- The dataset, provided by a Russian company 1C contains daily historical sales data of different shops and items sold.
- The objective of this project is to predict total sales for each item and store in the next month with the given Time-Series dataset.
- As a baseline method, decision tree classifier and regression on trees were implemented.
- Finally, a more sophisticated model of grouping sales data for each shop individually was devised which led to performing linear regression on individual shop and Item combination model.

Pseudo Code


Problem
Predict total sales for each item and store in the next month with the given Time-Series dataset
Aggregate
Aggregate sale count by month for each item in a shop.
Prepare
Prepare the data for analysis by generating the missing data points. (i.e each item in a shop should have 33 data points, where we replace '0' for missing values in the training dataset)
Perform
Perform analysis and test different prediction model on the prepared dataset.
Determine
Determine the best model to predict the sale count for next month for a given item in a shop.
Visualize
Visualize the actual and predicted values.
Create
Create a csv file with the predicted data for the given shop and items in the test dataset.

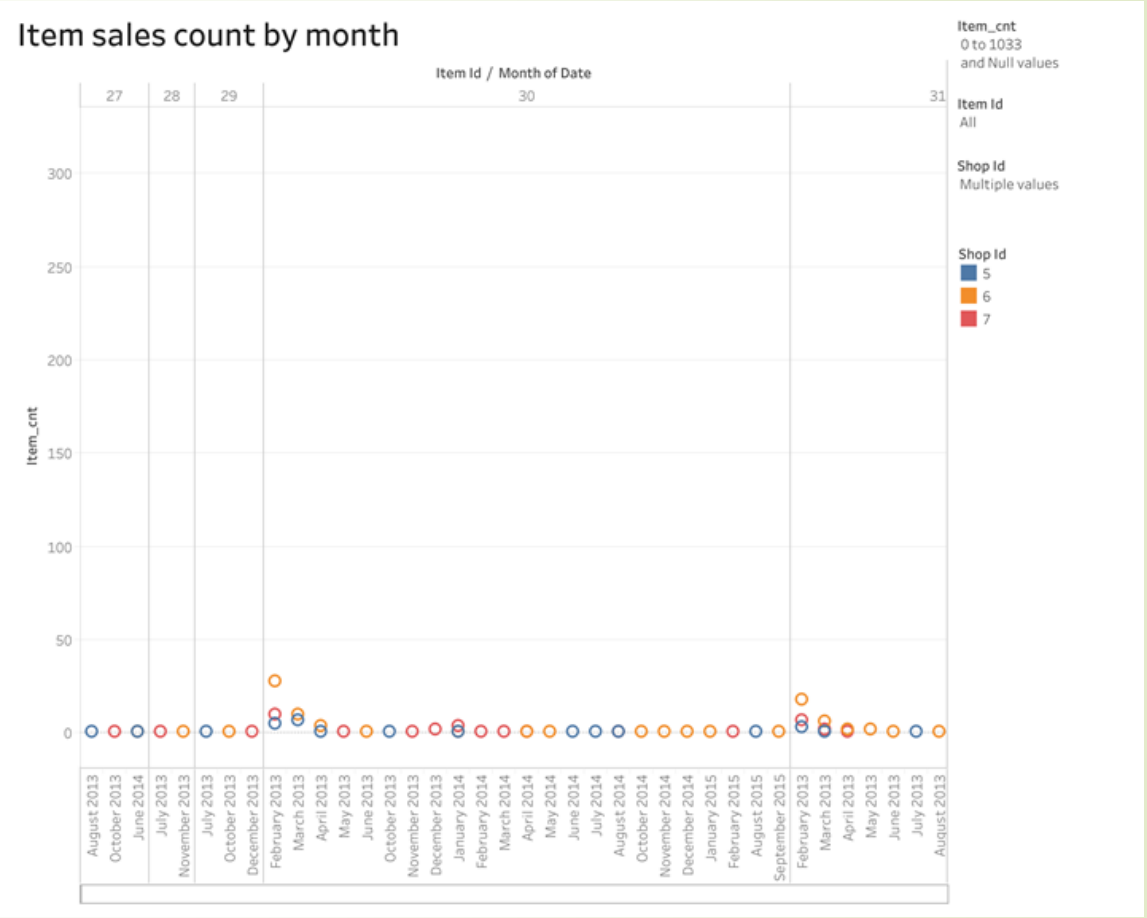
Exploratory Data Analysis

Variables of Interest

1. Shop_ID (Shop Identifier)
2. Item_ID (Item Identifier)
3. Date_Block_Num (Month Identifier)
4. Item_Cnt_Day (Item sale count)

On Aggregating the item sale count by month for a given shop, we produce the new aggregated metric **Item_Cnt.**

Click here to view the Tableau Report



Technical Approach

- Method 1: Decision Tree Classifier
- Method 2: Regression on Trees
- Method 3: Designing Linear Regression Model

METHOD 1: DECISION TREE CLASSIFIER

- The data was grouped by items sold in each shop and for each item in the shop
- Using sklearn, a decision tree model was trained on the training data provided to predict the item count sold for each shop and each item in the testing data provided.
- The model has predicted item count for the next month, date block number 34 and given shop and item ID

	shop_id	date_block_num	item_id	item_cnt
0	5	34	5037	1.0
1	5	34	5320	3.0
2	5	34	5233	1.0
3	5	34	5232	1.0
4	5	34	5268	1.0
5	5	34	5039	1.0
6	5	34	5041	2.0
7	5	34	5046	1.0
8	5	34	5319	3.0
9	5	34	5003	1.0

METHOD 2: REGRESSION ON TREES

- The data prepared for the decision tree classifier was used to implement regression on trees
- Using sklearn, a decision tree regressor was trained on the training data to predict the item count sold for each shop and each item in the testing data provided.
- The decision tree regressor model has predicted item count for the next month, date block number 34 and given shop and item ID.

	shop_id	date_block_num	item_id	item_cnt
0	5	34	5037	1.0
1	5	34	5320	1.0
2	5	34	5233	1.0
3	5	34	5232	1.0
4	5	34	5268	1.0
5	5	34	5039	1.0
6	5	34	5041	2.0
7	5	34	5046	1.0
8	5	34	5319	1.0
9	5	34	5003	1.0

METHOD 3: DESIGNING A LINEAR REGRESSION MODEL

- In this approach, combination of individual shop and items in the shop separately are treated as a model
- A prototype was run only for shop ID 5 considering the data size.
- To fill the gap of missing data of an item for any of the months between the first and the last, a data point of item count sold was added with the value of zero
- Using statsmodel, a linear regression model was trained on the prepared data

5_1830

OLS Regression Results

Dep. Variable:	item_cnt_agg	R-squared:	0.522
Model:	OLS	Adj. R-squared:	0.507
Method:	Least Squares	F-statistic:	34.94
Date:	Fri, 12 Apr 2019	Prob (F-statistic):	1.41e-06
Time:	12:50:19	Log-Likelihood:	-87.782
No. Observations:	34	AIC:	179.6
Df Residuals:	32	BIC:	182.6
Df Model:	1		
Covariance Type:	nonrobust		

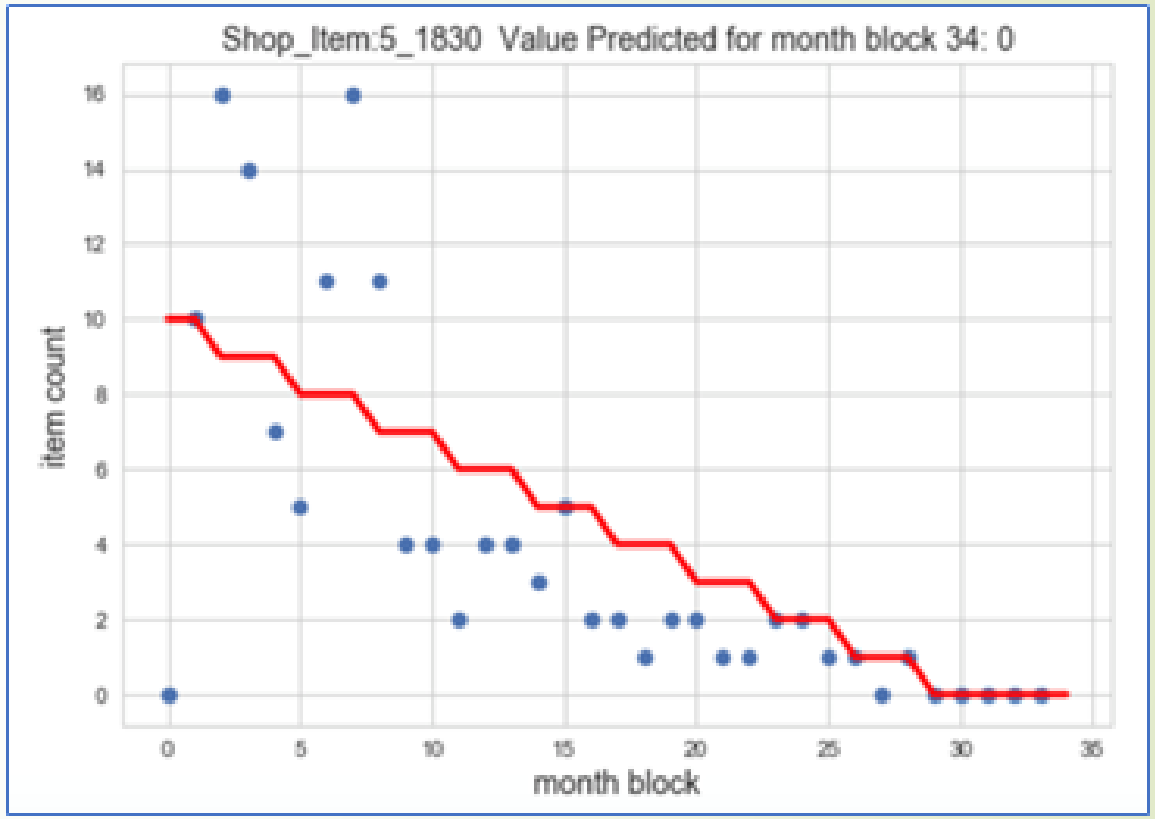
	coef	std err	t	P>|t|	[0.025	0.975]
const	9.5630	1.107	8.642	0.000	7.309	11.817
date_block_num	-0.3407	0.058	-5.911	0.000	-0.458	-0.223

Omnibus:	6.338	Durbin-Watson:	1.040
Prob(Omnibus):	0.042	Jarque-Bera (JB):	7.731
Skew:	0.277	Prob(JB):	0.0210
Kurtosis:	5.269	Cond. No.	37.6

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

EXAMPLE OF PREDICTED OUTPUT FOR ONE MODEL

- There were few shops and items ids which were present in testing data but not in training data for which our model predicted NA.
- The prediction for shop ID 5 and item ID 1830 for the next month, date block number 34 can be seen below. The model has predicted 0 as the item count
- Blue dots are the actual sales and red line is the predicted values from our model.



CONCLUSION

- Designed three approaches to predict future sales
- first two with decision tree classification and decision tree regression showing an example of supervised learning method.
- To devise a better approach for a strong model training mechanism, moved to the third and final approach of implementing "linear regression" by creating a regression model for every shop_item combination.
- Linear Regression approach is better than others because it formulates a strong relation between item and consecutive months numbers for a given shop.
- As a part of deliverables, we also generated a csv file with the predicted data for the given shop and items in the test dataset.

THANK YOU