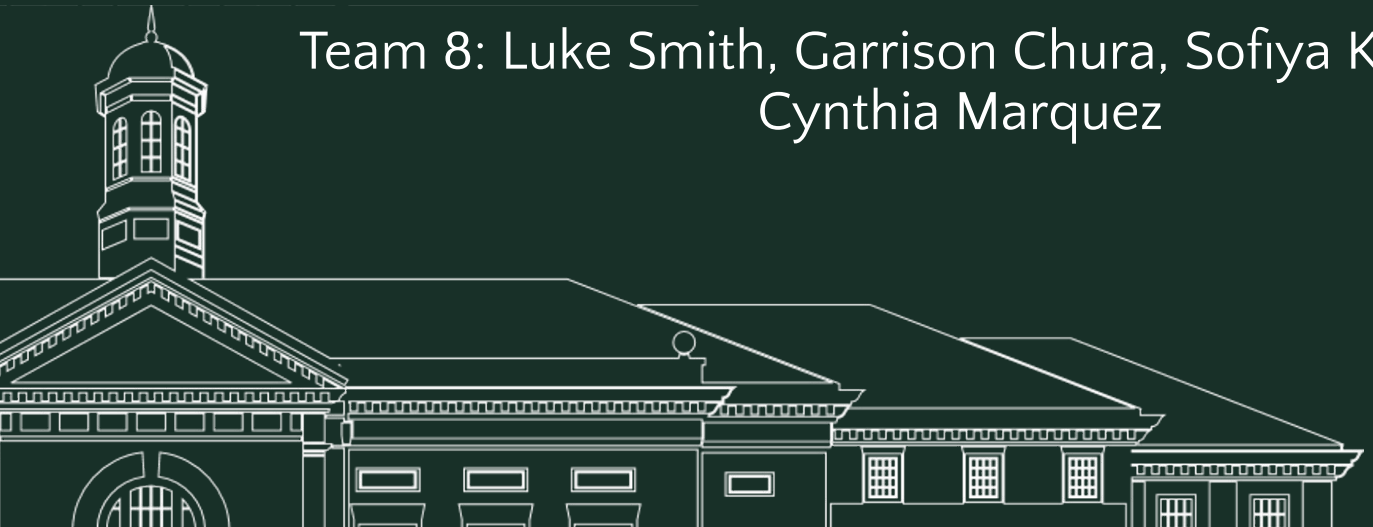
Raymond A. Mason
School of Business
WILLIAM & MARY

Competition Research: Predict Future Sales

Team 8: Luke Smith, Garrison Chura, Sofiya Kuzina, and Cynthia Marquez



Overview/Problem Statement

- Kaggle dataset
- Time-series
- Daily sales data
- 1C Company

- **<u>Objective</u>: Identify the sales metrics that are most important to predict average sales of a particular month**

Data details

File descriptions:

sales_train.csv – the training set. Daily historical data January 2013 – October 2015

test.csv – the test set. Use this to forecast sales for November 2015

sample_submission.csv – a sample submission file in the correct format

items.csv – supplemental information about items/products

item_categories.csv – supplemental information about items categories

shops.csv– supplemental information about shops

Data fields:

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. Predict a monthly amount of this measure

item_price – current price of an item

date – date in format dd/mm/yyyy

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

Critiques

- Common critiques from the following published notebooks:
  - “Predict Future sales R (shop wise model)”
    - Predict Future sales R (shop wise model) | Kaggle
  - “Prediction for Future Sales”
    - 📈 Prediction for Future Sales | Kaggle

Critiques

- Use of SQL statements

Predict Future sales R (shop wise model) | Kaggle

```
Test_prep = sqldf("SELECT a.item_id, max(a.item_price) as item_price
FROM sales_train a
inner JOIN (
SELECT item_id, MAX(date) as date,MAX(shop_id)
FROM sales_train
GROUP BY item_id
) b
ON a.item_id = b.item_id
and
a.date = b.date
group by a.item_id")


Test_data = sqldf("select t1.id,t1.shop_id,t1.item_id,t3.item_category_id, 34 as date_block_num,t2.item_price from test t1
left join
Test_prep t2
on
t1.item_id = t2.item_id
left join
items t3
on t1.item_id = t3.item_id")
```

Critiques

- Does linear regression model real-world scenarios?

```
model <- lm(formula = item_cnt_month ~ . , data = train_data_df)

result = predict(model, test_data_df_predict)
submission =  data.frame(ID = test_data_df$id,
                         item_cnt_month = result)
submission_results = union(submission_results,submission)
```

Predict Future sales R (shop wise model) | Kaggle

```
model<-lm(sold_qties~shop_id + item_id + item_price + item_category_id, data=train)

summary(model)
```

📈 Prediction for Future Sales | Kaggle

Critiques

- Other Considerations:
  - Comparison of models to determine best performance
    - Statistical tests and metrics
  - Data cleansing
    - Handling missing values
  - Explainability/Analysis

Proposed Solution & Reproducibility

Github and References

https://github.com/cvmarquezt/ML2

https://www.kaggle.com/competitions/competitive-data-science-predict-future-sales/overview

https://www.kaggle.com/code/manojlukhi/predict

https://www.kaggle.com/code/saikat026/prediction-for-future-sales