Dillard's Profit Prediction

— MSiA 400 Final Project Team 12

Team Members:

Alejandra Lelo De Larrea Ibarra

Yi Chen

Yiqing Cheng

Xin (Susie) Shu

Pictures taken from 293 Dillards Images, Stock Photos & Vectors | Shutterstock, Wealth Management, Dillards Website

Content

- **Executive Summary**
- **Importance of Forecasting Profits**
- **Data**
- **Modeling**
- **ROI Analysis**
- **Conclusion**

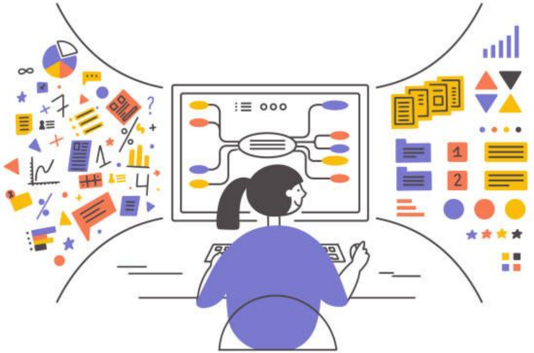
Executive Summary

About Dillard’s Inc.

- One of the largest fashion retailers in the US
- Cares about profit forecasting
- Use SARIMAX model to predict monthly profits*

Our Roles and Plan

- Data scientist hired by Dillard’s Inc. to develop ML model and increase profit prediction accuracy
- Predict profits with data on national level; augmented with macroeconomic indicators with four models: SARIMAX(Baseline) Facebook Prophet, Lasso Regression, XGBoost



Our Results

- Lasso is the model that performs best
- Forecast the profit of the last month in the sample, August 2005, and find that around 91%(66%) of the variability can be explained by the model for state level (national) aggregates

Pictures taken from Wealth Management, https://www.istockphoto.com/illustrations/data-scientistc

This is an assumption of our project. The SARIMAX model will be trained in modeling section

Content

Executive Summary

Importance of Forecasting Profits

Data

Modeling

ROI Analysis

Conclusion

Importance of Forecasting Profits

Estimation of Profits

↓

Estimation of Future Cash Flow

↓

Long-term Strategy

Business Expansion

Marketing Strategy

Goal

Forecast the daily profit with internal daily transactions data provided by the company.

Tools

SARIMAX

FB Prophet

Lasso

XGBoost

SARIMAX is the baseline model, and we compare predictions from other three models with SARIMAX.

Content

Executive Summary

Importance of Forecasting Profits

Data

Modeling

ROI Analysis

Conclusion

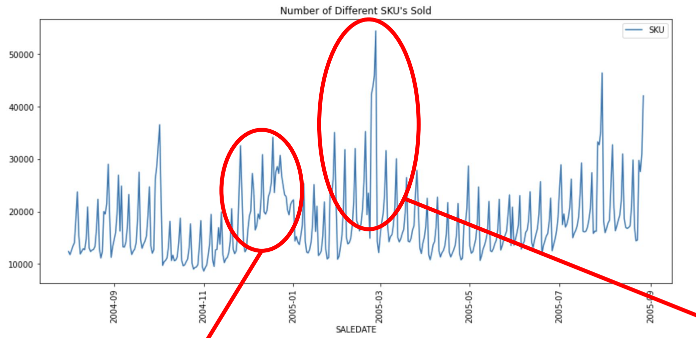
Data

The data provided by the company spans from August 1st, 2004 to August 27th, 2005 with information:

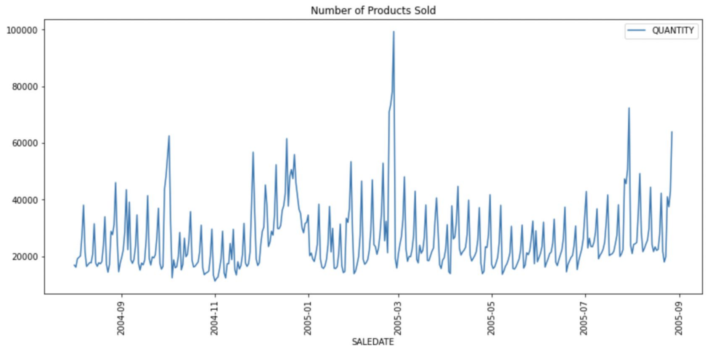
- **Store Transaction Data from each transaction at each Dillard's store**
- **Characteristics of each product**
- **Cost and retail price of every product**
- **Location of stores**

Exploratory Data Analysis

Number of Different SKU's Sold During Periods



Number of Product Sold During Periods



The time series plot of the total number of products sold each day has some **expected peaks**. For example, around **Thanksgiving**; however, we found that there was an **unexpected peak on February 26th, 2005** which, after researching, we found correspond to the "President's sale day" one of the days of the year with the highest discounts

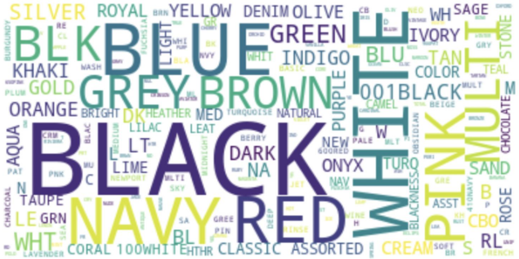
Data

The data provided by the company spans from August 1st, 2004 to August 27th, 2005 with information:

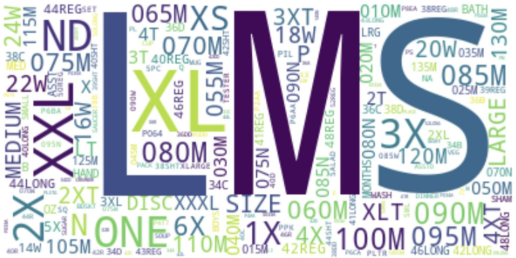
- Store Transaction Data from each transaction at each Dillard’s store
- Characteristics of each product
- Cost and retail price of every product
- Location of stores

Exploratory Data Analysis

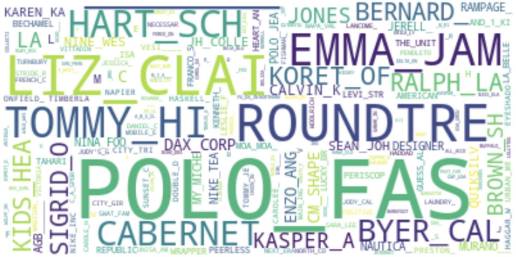
The most common transacted color



The most common transacted size



The most common transacted brand



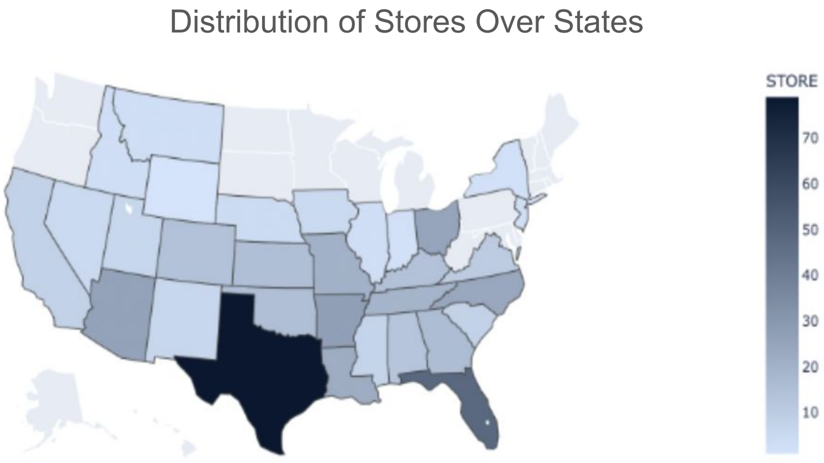
We can detect that the **original data is messy** since there are different interpretations of the same term. For example, for the black color we can find "black", "blk", "blac", "001black" and all of these must be marked as "black". So **further cleaning is needed** if this variable will be used in the modelling part.

Data

The data provided by the company spans from August 1st, 2004 to August 27th, 2005 with information:

- **Store Transaction Data from each transaction at each Dillard's store**
- **Characteristics of each product**
- **Cost and retail price of every product**
- **Location of stores**

Exploratory Data Analysis



The number of stores over states varies. South states seems to have more stores. It's possible that the profits also varies among states.

Other Findings

There is an **original price** and also a **retail price** for the products, we assume that original price is the first price (or suggested price) of the article and that retail price is the actual price paid by the client in the transaction, thus some articles were sold at discount

Data

The data provided by the company spans from August 1st, 2004 to August 27th, 2005 with information:

- **Store Transaction Data from each transaction at each Dillard’s store**
- **Characteristics of each product**
- **Cost and retail price of every product**
- **Location of stores**

Feature Engineering

Calculate Daily Profit

$\Pi kt = pkt - ckt$ where pkt is the retail price for product k at day t

Create Sale - Correlated Features

1. Number of Stores Selling Each Product; 2. Number of Different Products Sold;
3. Average Quantity Sold Per Transaction; 4. Average Cost Per Transaction;
5. Max, Min, Avg Prices and Costs Per Transaction;
6. Number of Vendors, Departments, Cities, States

Lagged Features

Features lags 1, 2, 3, 4, 5, 6, 7, 14, 21, 28 days to catch seasonality and prevent data leakage

Time-related Features

Year, Quarter, and Week of the year, are generated to capture the time effects for general machine learning model

Content

Executive Summary

Importance of Forecasting Profits

Data

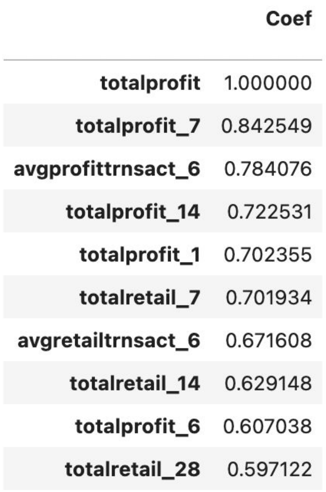
Modeling

ROI Analysis

Conclusion

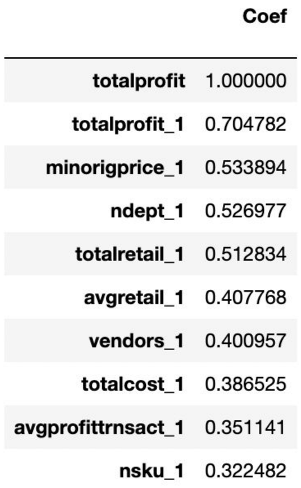
Modeling – Feature Selection

- For ARIMA and Facebook Prophet, we use correlation coefficient to select features to be included
- For Lasso and XGBoost we input all variables and let the data speak for itself



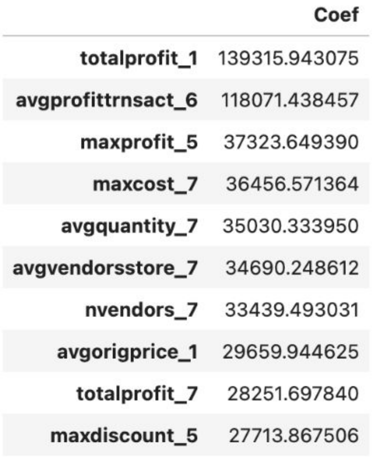
	Coef
totalprofit	1.000000
totalprofit_7	0.842549
avgprofittrnsact_6	0.784076
totalprofit_14	0.722531
totalprofit_1	0.702355
totalretail_7	0.701934
avgretailtrnsact_6	0.671608
totalretail_14	0.629148
totalprofit_6	0.607038
totalretail_28	0.597122

ARIMA



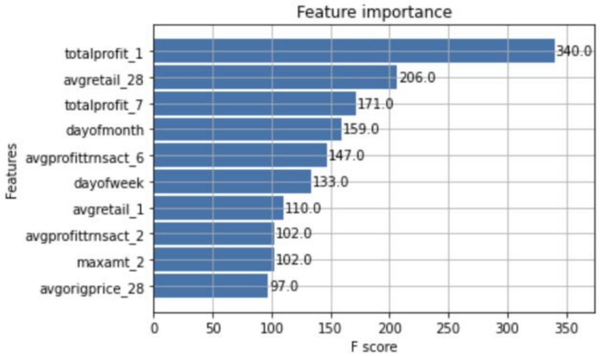
	Coef
totalprofit	1.000000
totalprofit_1	0.704782
minorigprice_1	0.533894
ndept_1	0.526977
totalretail_1	0.512834
avgretail_1	0.407768
vendors_1	0.400957
totalcost_1	0.386525
avgprofittrnsact_1	0.351141
nsku_1	0.322482

Facebook Prophet



	Coef
totalprofit_1	139315.943075
avgprofittrnsact_6	118071.438457
maxprofit_5	37323.649390
maxcost_7	36456.571364
avgquantity_7	35030.333950
avgvendorsstore_7	34690.248612
nvendors_7	33439.493031
avgorigprice_1	29659.944625
totalprofit_7	28251.697840
maxdiscount_5	27713.867506

Lasso



XGBoost

Modeling – Model Introduction

Time Series Models

Facebook Prophet

Good at modeling time series that have multiple seasonality

SARIMAX

Consider exogenous features and seasonality

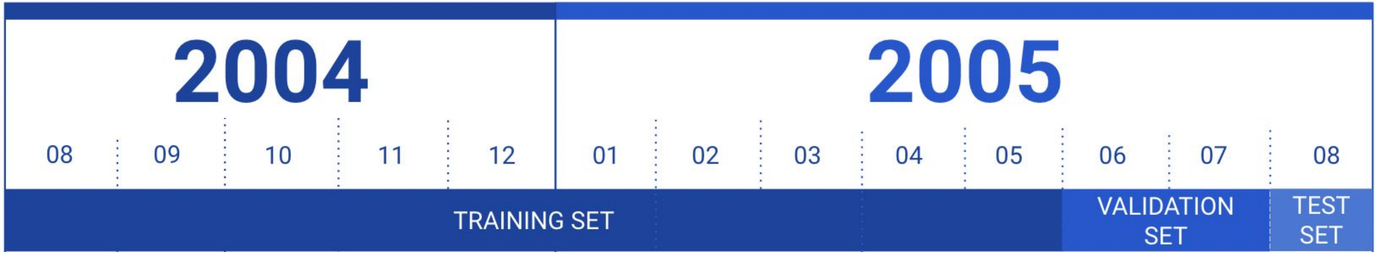
General Machine Learning Models

Lasso Regression

Conduct feature selection based on linear regression

XGBoost

Variate gradient boosting model



2004

2005

08 09 10 11 12 01 02 03 04 05 06 07 08

TRAINING SET

VALIDATION SET

TEST SET

Modeling – Model Performance

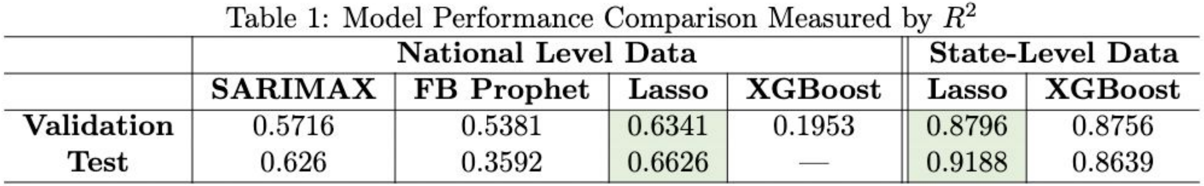
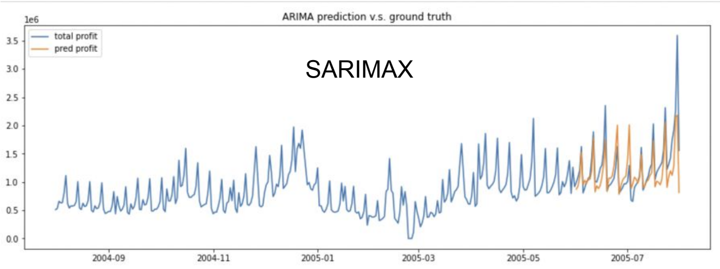


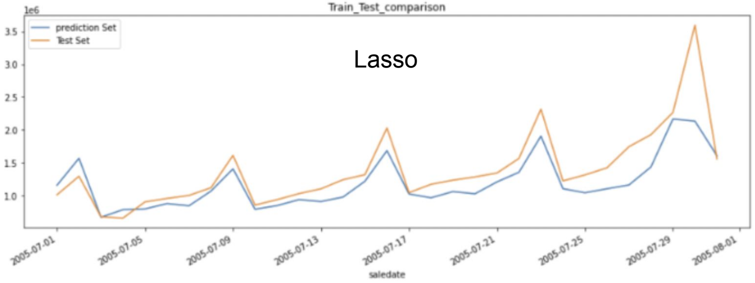
Table 1: Model Performance Comparison Measured by R^2

	National Level Data				State-Level Data	
	SARIMAX	**FB Prophet**	**Lasso**	**XGBoost**	**Lasso**	**XGBoost**
Validation	0.5716	0.5381	0.6341	0.1953	0.8796	0.8756
Test	0.626	0.3592	0.6626	—	0.9188	0.8639

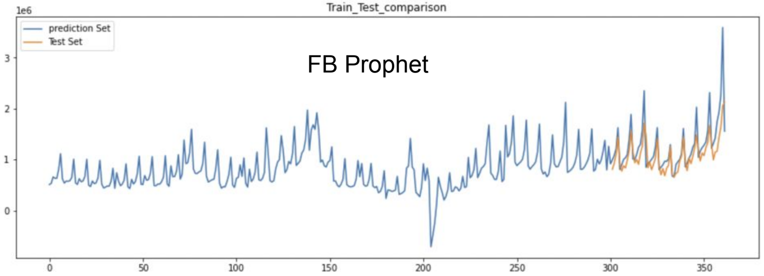
- Lasso performs the best on both national and state level
- Models on state-level data performs better than national level data



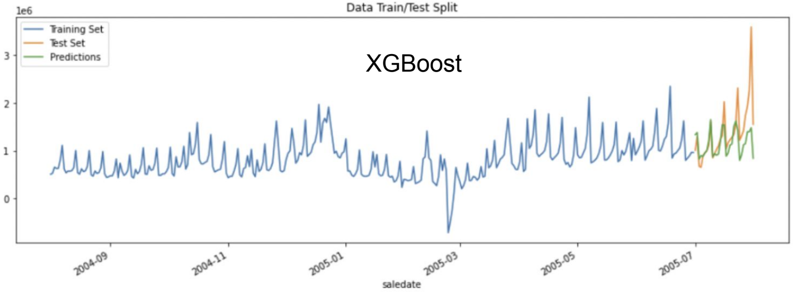
SARIMAX



Lasso



FB Prophet



XGBoost

Modeling – Prediction Results

- SARIMAX, Prophet, Lasso and XGBoost on national level data give **underestimation** on profit in Aug

- Lasso and XGBoost on state-level data give **overestimation** on profit in Aug

- For forecasting profits it is better to have underestimation. Despite the R2, we recommend the Lasso model on national level data when estimating profits.

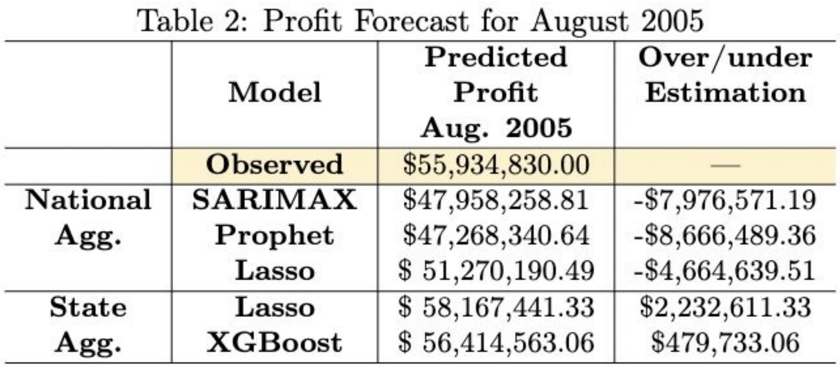


Table 2: Profit Forecast for August 2005

	Model	Predicted Profit Aug. 2005	Over/under Estimation
	Observed	$55,934,830.00	—
National Agg.	**SARIMAX**	$47,958,258.81	-$7,976,571.19
	Prophet	$47,268,340.64	-$8,666,489.36
	Lasso	$ 51,270,190.49	-$4,664,639.51
State Agg.	**Lasso**	$ 58,167,441.33	$2,232,611.33
	XGBoost	$ 56,414,563.06	$479,733.06

Content

Executive Summary

Importance of Forecasting Profits

Data

Modeling

ROI Analysis

Conclusion

ROI Analysis

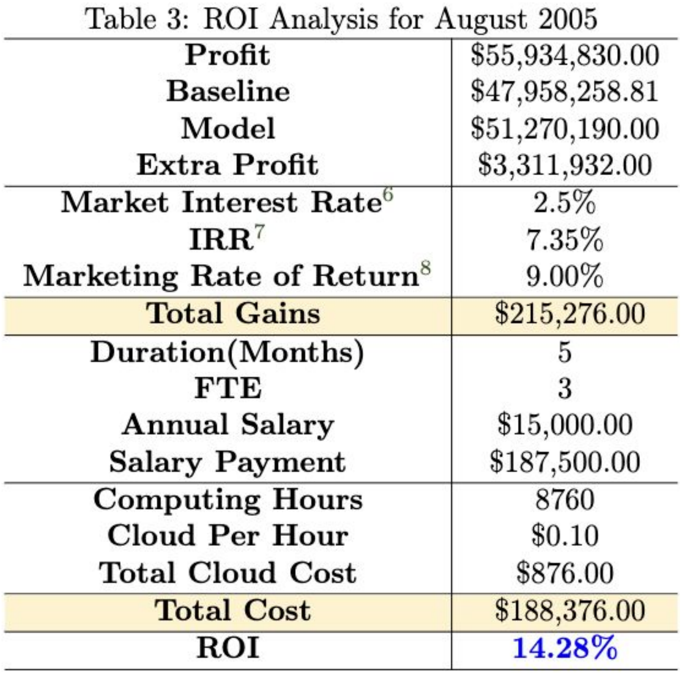


Table 3: ROI Analysis for August 2005

Profit	$55,934,830.00
Baseline	$47,958,258.81
Model	$51,270,190.00
Extra Profit	$3,311,932.00
Market Interest Rate[6]	2.5%
IRR[7]	7.35%
Marketing Rate of Return[8]	9.00%
Total Gains	$215,276.00
Duration(Months)	5
FTE	3
Annual Salary	$15,000.00
Salary Payment	$187,500.00
Computing Hours	8760
Cloud Per Hour	$0.10
Total Cloud Cost	$876.00
Total Cost	$188,376.00
ROI	14.28%

- **Assumption**
  - All cash can be reinvested in marketing project earning rate of return at 9.00%
  - Dillard will deposit profits into bank if there is no profitable investment project
  - Interest rate for deposit into bank: 2.5%
  - IRR: 7.35%
  - Need 3 data scientists working for 5 month
  - Cloud based model infrastructure

- **Calculation of ROI**
  - **Total extra gains from lasso model**

    Extra Profit x (1.09- 1.025)

  - **Total cost with prediction model**

    Salary Payment + Total Cloud Cost

  - **ROI**

    (Gains - Cost) / Cost

[6] The Market Interest Rate is the average Federal Funds Rate between 2004 and 2005

[7] The Internal Rate of Return for Dillard's is estimated from Dillard's Weighted Average Cost of Capital since a retailer's cost of capital is used as the required rate of return.

[8] The Required Rate of Return of marketing strategy is estimated from Nielson's Maximize the Return on your Advertising Spend.

Content

Executive Summary

Importance of Forecasting Profits

Data

Modeling

ROI Analysis

Conclusion

Conclusion

Modeling Conclusion

- On the national level, Lasso performs the best for forecasting Dillard’s profit with an R^2 of 0.6341, followed by SARIMAX with an R^2 of 0.5716
- On the state level, Lasso remains the best model with an R^2 of 0.8796 on the validation set. As for the prediction of total profits in August 2005, the model that best performs is Lasso for state-level data
- Models trained on nationwide aggregated underestimate the observed profit and models trained on statewide aggregated data overestimate this measure
- But for forecasting profits, it is better to have a more conservative estimate (i.e. underestimation), it is our advise that the Lasso model trained with national level data be the one used when estimating profits

ROI Conclusion

- The informed investments with better forecasts can generate extra revenue with ROI of 14.28%

Future Development

- Include extra external features: stock data of Dillard’s, measures of risk(e.g. inflation on producers prices), Dillard’s inventory data
- Increase sample size: It could be helpful to obtain more history on transactions to include yearly patterns