



Office Supply Store

By Sahar Sami



Contents

- Background
- Objectives
- Approach
- Analysis Results
- Recommendations
- Appendix



Background

- Office supply store has tested a telemarketing campaign on 16,000 customers
- Client wants to use findings from the campaign to determine which customers to target for future campaigns



Objectives

- Profile the customers that responded to the campaign to make the company aware of who its typical customer is
- Develop models that will allow the company to use the results of the campaign to target future, similar campaigns
- Show the financial value of your models



Approach



Data Analysis

- Exploratory data analysis on provided dataset
- Clean dataset and perform feature engineering



Model Creation

- Extra Trees Classification to determine probability of sale
- Random Forest Regression to predict transaction size
- Calculate estimated profit using model results



Financial Analysis

- Determine which characteristics produce the highest profit
- Recommend what types of customers to target in future campaigns

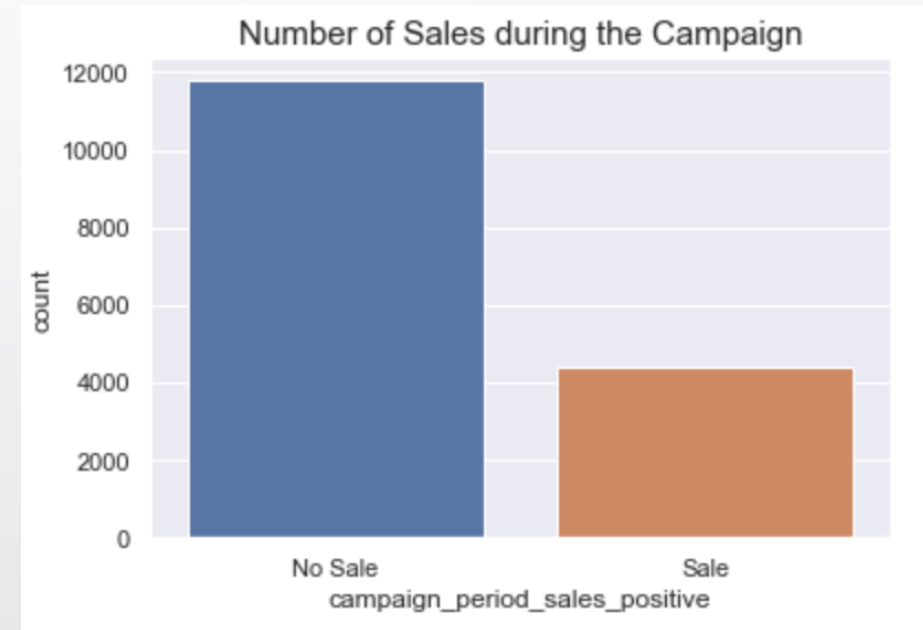


EDA & Preprocessing

- Data features
 - Sales made during the campaign period
 - Historical sale amount
 - Date of first purchase
 - Prior transaction information
 - Prohibited communication channels
 - Number of employees
 - Primary language
- Some missing values
 - Removed observations with mostly missing variables
 - Created 'unknown' category for categorical variables with missing values
- Removed observation with outlier in number of past transactions

EDA & Preprocessing

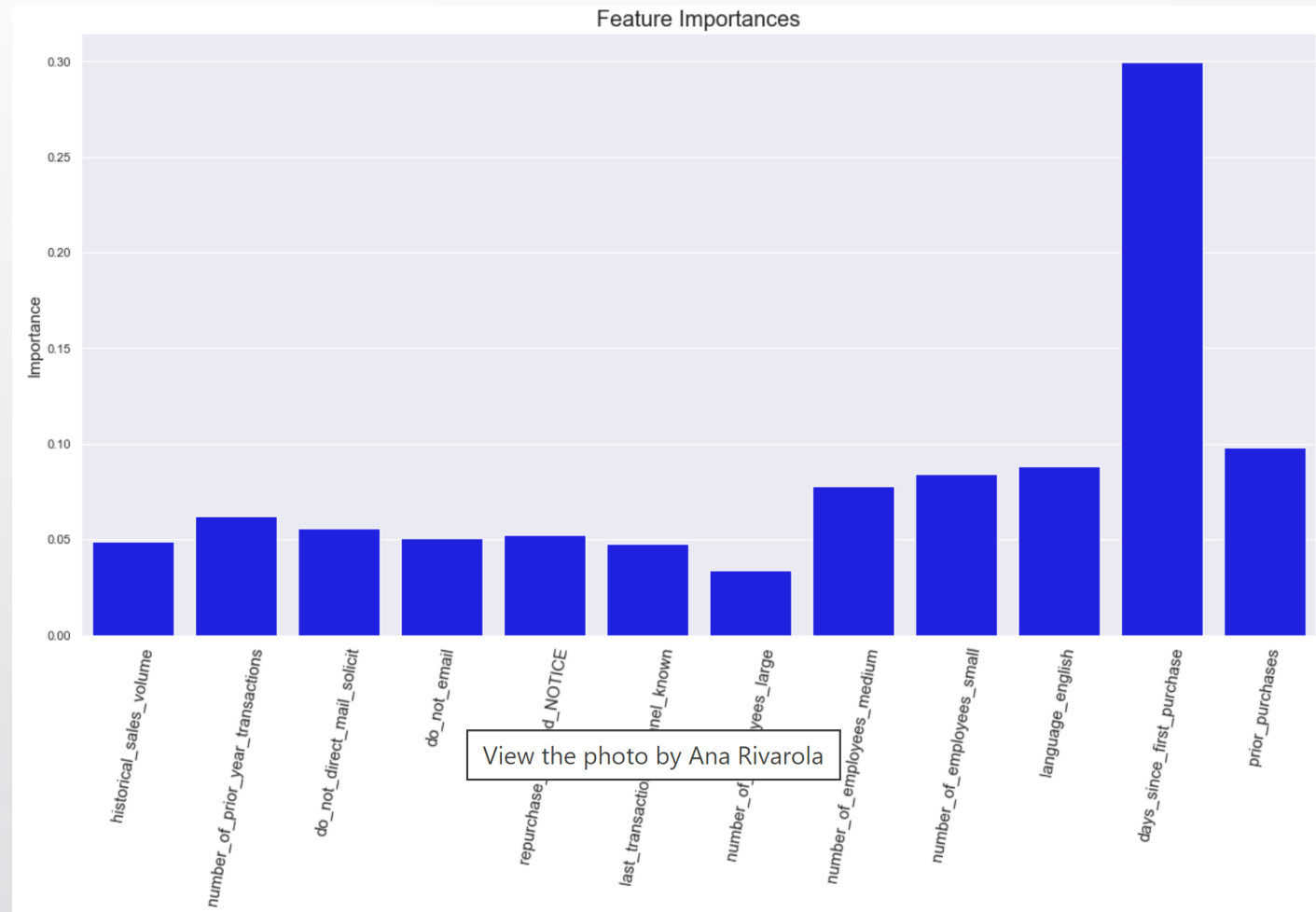
- Created a binary variable to indicate positive campaign period sales: most companies did not make a purchase
- Created a new feature representing the number of days since a company's first purchase
- Created dummy variables for categorical features including number of employees and last repurchase method



About 27% of companies made a positive transaction during the campaign

Classification – How likely is a sale?

- Model trained on 80% of existing business customer data, oversampled for even classes
- Extra Trees model performed the best out of 10 models with 0.985 AUC score and 94% accuracy
- Most important features: days since first purchase, language is/is not English, number of prior purchases

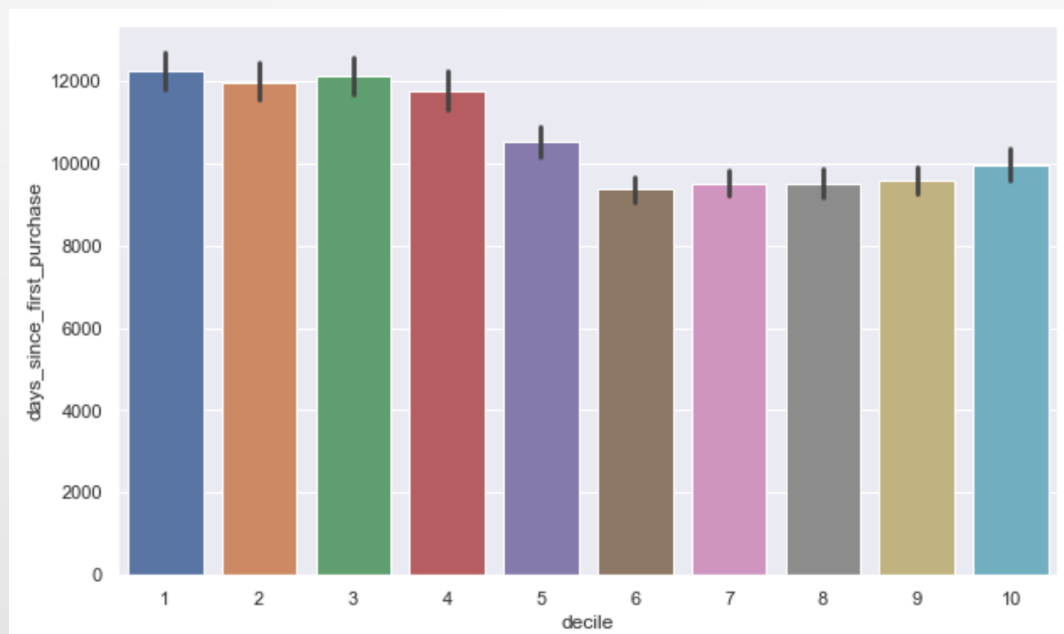


Classification Gain Chart

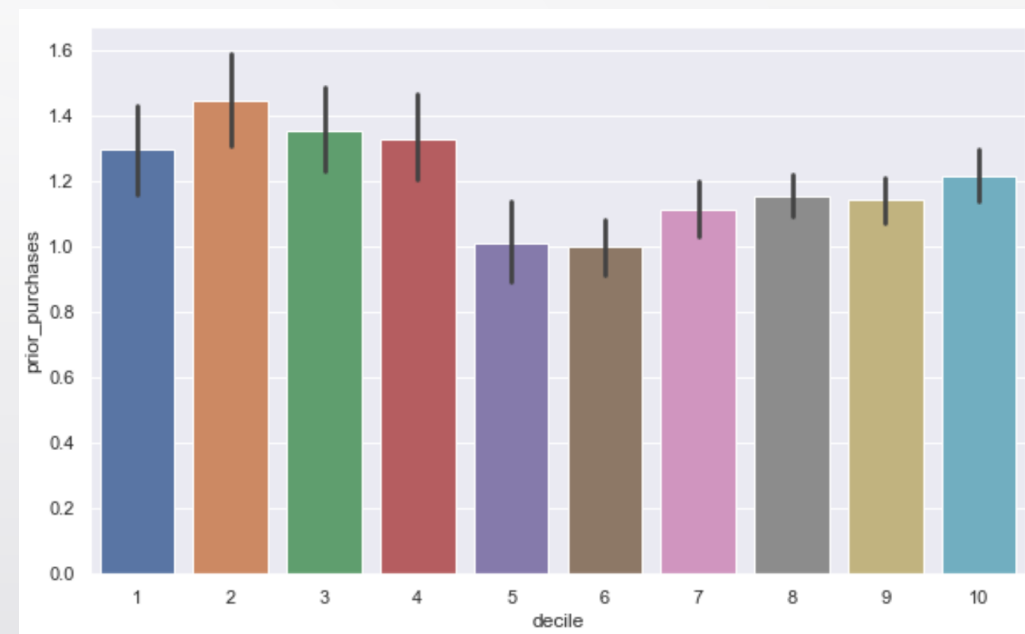
- Deciles 1, 2, 3, and 4 have the highest number of customers with positive sales
- Decile 4 has the highest score of separation between positive and zero sales

Decile	Number of Pos. Sales	Total Customers	Default_Rate	Default %	KS Statistic	Max KS Statistic	default_cum %	Gain	Lift
1	462	471	98.09%	19.95%	19.57		19.95	19.95	1.99
2	471	472	99.79%	20.34%	39.87		40.28	40.28	2.01
3	467	472	98.94%	20.16%	59.82		60.45	60.45	2.02
4	465	472	98.52%	20.08%	79.61	*****	80.53	80.53	2.01
5	192	472	40.68%	8.29%	76.24		88.82	88.82	1.78
6	27	472	5.72%	1.17%	58.87		89.98	89.98	1.50
7	32	472	6.78%	1.38%	41.93		91.36	91.36	1.31
8	37	472	7.84%	1.60%	25.41		92.96	92.96	1.16
9	60	472	12.71%	2.59%	10.84		95.55	95.55	1.06
10	103	470	21.91%	4.45%	0		100	100	1

Decile Analysis



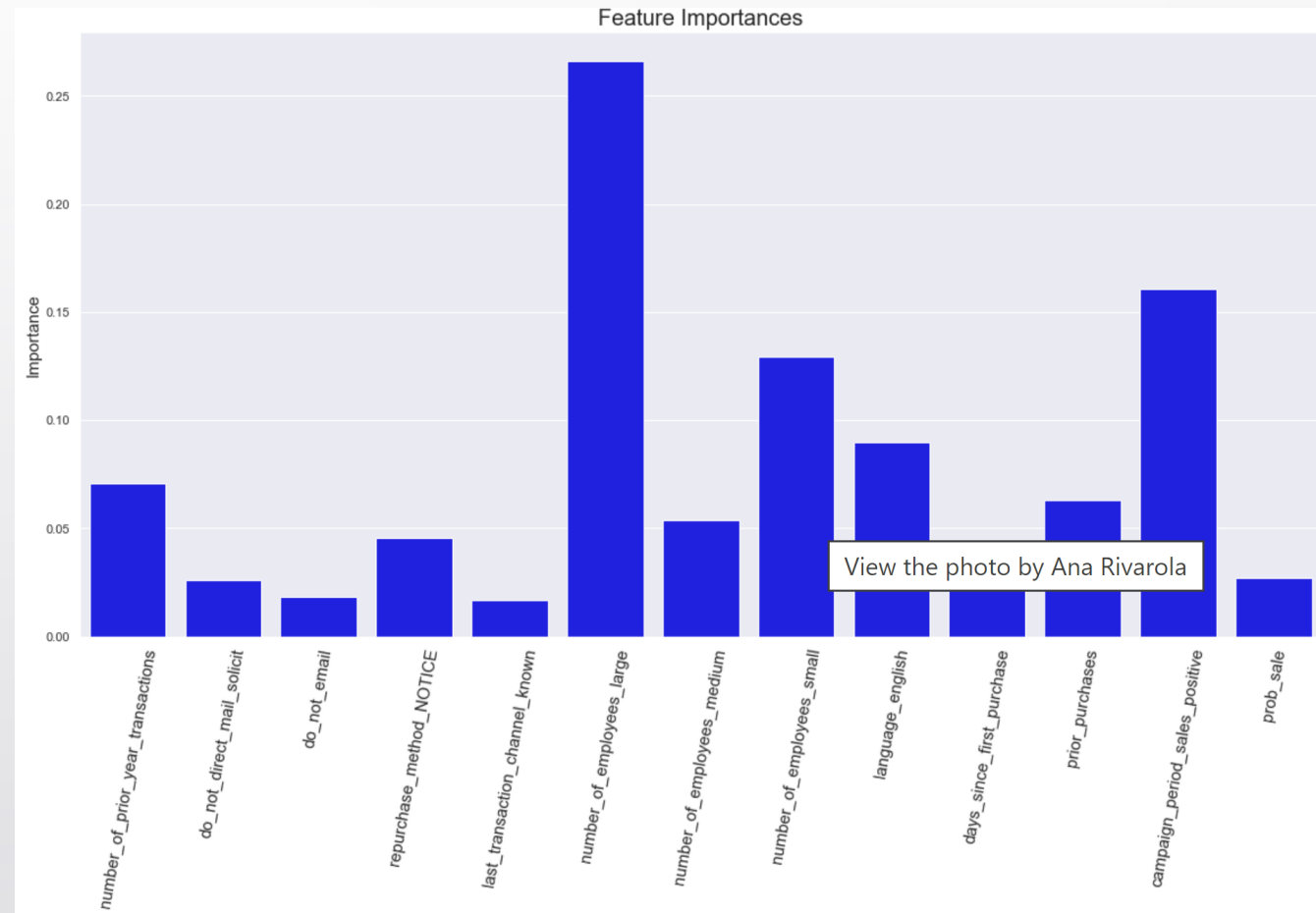
Deciles with a higher number of positive observations have a longer average tenure



Deciles with a higher number of positive observations have a higher average number of prior purchased items

Regression – How large will the transaction be?

- Model trained on 80% of the same customer data
- Random Forest Regression performed the best out of 10 models with the smallest mean squared error
- Most important features are whether the company has a large number of employees and if the campaign period sales were positive

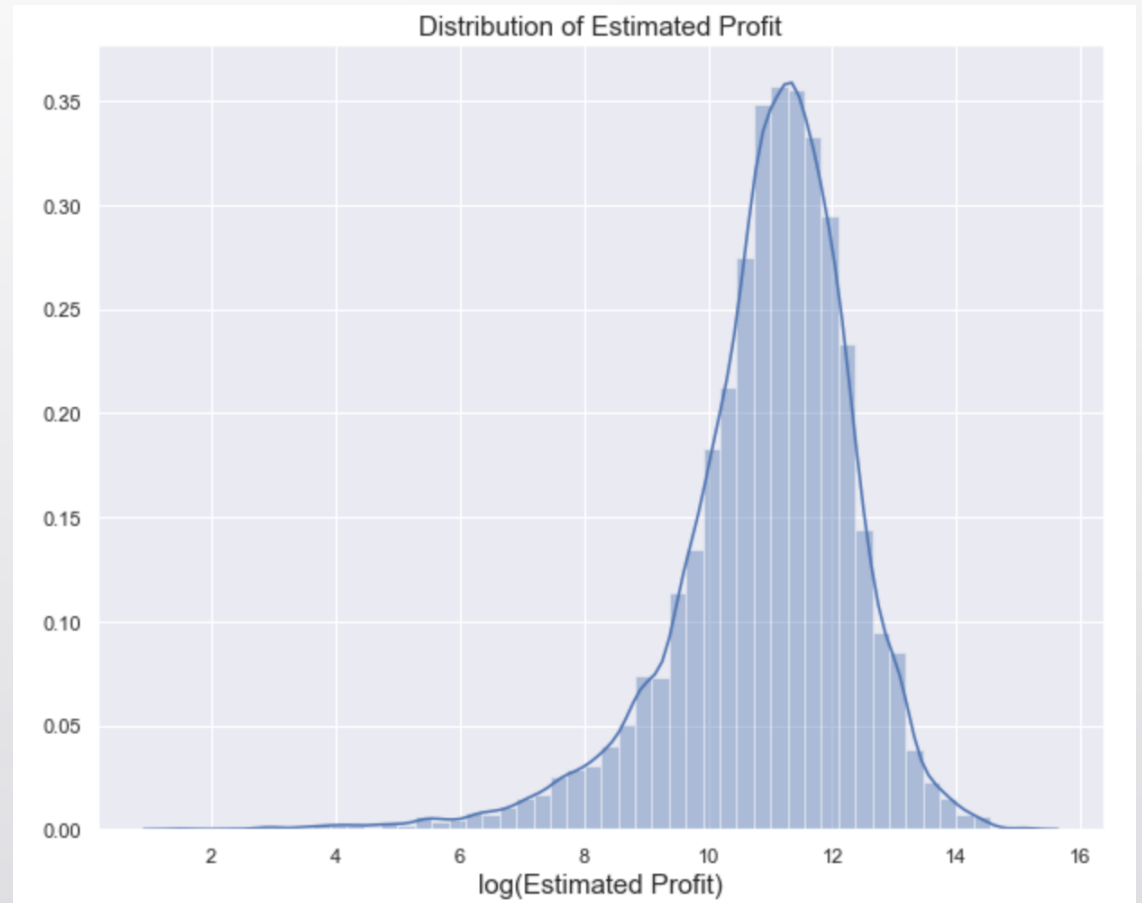


Profitability Computation

- Gross margin on sales: 22%
- Campaign cost: \$45.65 per business contacted
- Transaction cost: \$8.40 per transaction

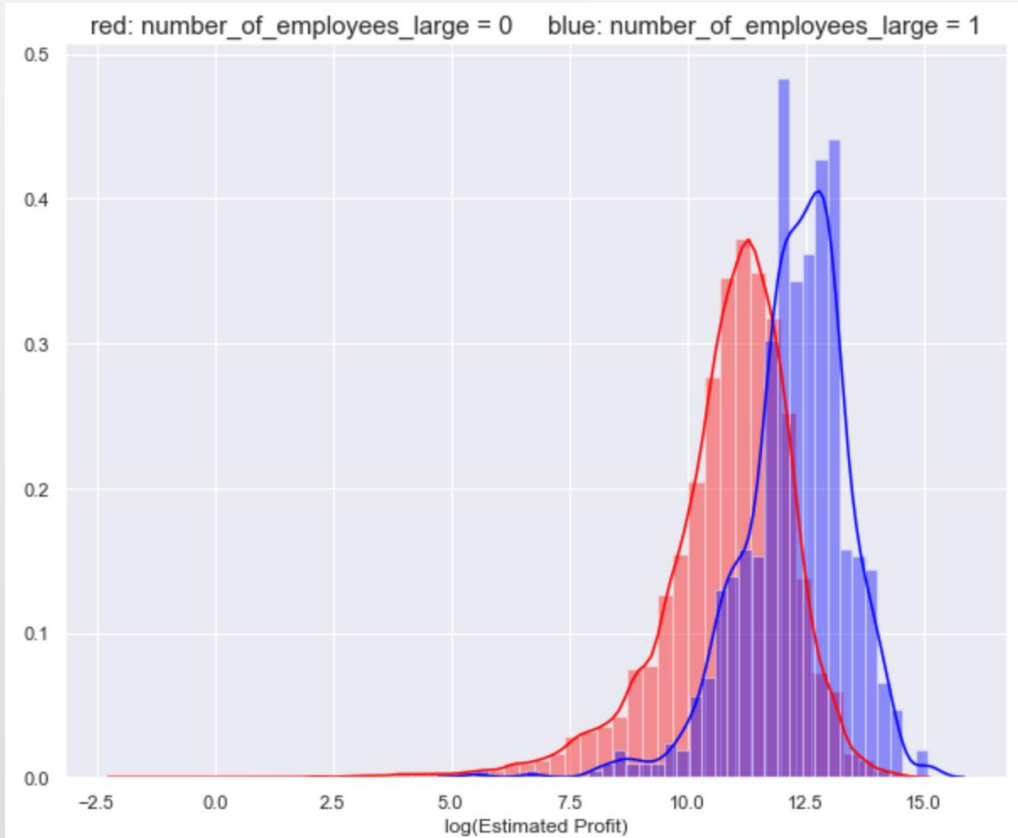
$$\begin{aligned} E(\text{Profit}) &= .22 * \text{Prob}(\text{Sale}) \\ &\quad * \text{Est}(\text{Transaction Size}) - \$8.40 \\ &\quad * \text{Prob}(\text{Sale}) - \$45.65 \end{aligned}$$

- Average Profit: \$658,991



Analysis of Important Features

Distribution of Estimated Profit



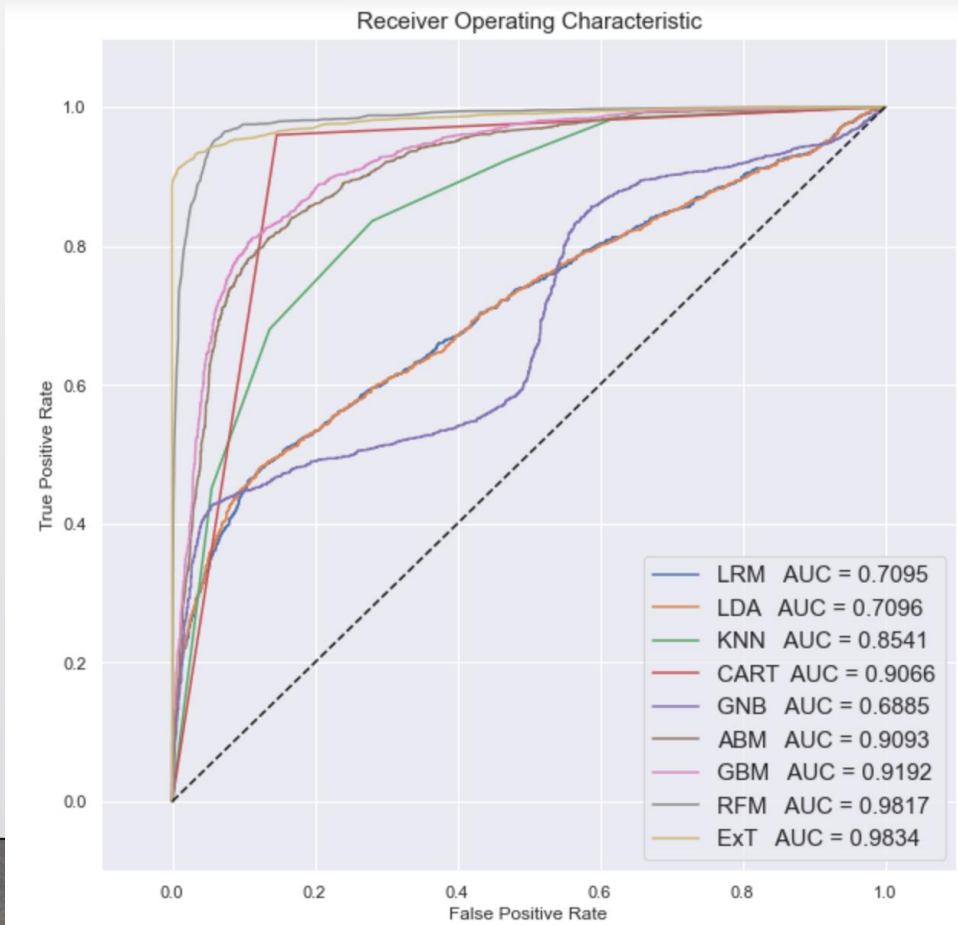
Large companies (101+ people) tend to have a higher estimated profit



Recommendations for Future Campaigns

- Target companies with a high tenure who purchased more products in the past
- Of those companies, target large companies (101+ employees)
- Research number of employees for companies where size is unknown
- Research the primary language of companies where it is unknown

Appendix: Area under ROC Curve for Classification



- Area under ROC Curve (AUC) metric was used to determine the best models
- Random Forest (RFM) and Extra Trees (ExT) were both tuned to find the best of the two models under the best hyperparameters