BANCO DE PORTUGAL
EUROSISTEMA

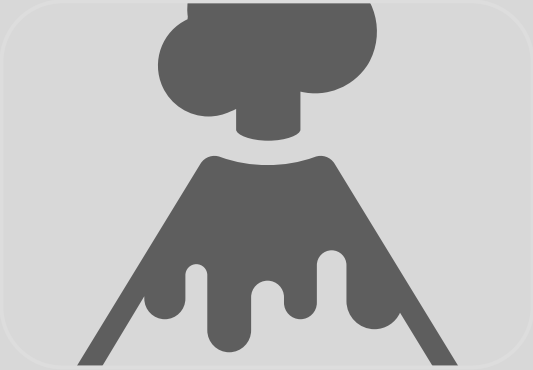
PREDICTING BANK TELEMARKETING SALES

Nick Gigliotti

Business Problem

Banco de Portugal hired me to develop an accurate **predictive model** to predict which customers are likely to invest in a term deposit as a result of telemarking.

Telemarketing is hard on both salespeople and customers.

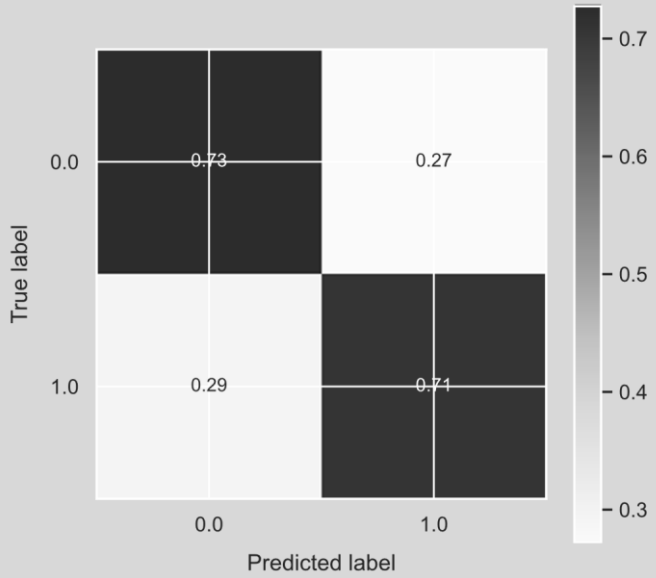


Connecting with the wrong customers results in **unpleasant conversations** and desperate marketing tactics.

Effectively connecting salespeople with high-potential customers reduces stress and increases profit.

Data and Methods

- Banco de Portugal Telemarketing Dataset
  - Publicly available on the UCI Machine Learning Repository
  - Collected from May 2008 to November 2010
  - Originally contains 21 features and 41k samples
  - Roughly two thirds categorical and one third numeric



- Methods
  - Logistic Regression
    - Relatively simple
    - Powerful
    - Widely used
  - Iterative Progress
    - Split the data into a 75% training set and 25% test set
    - Start with a dummy model and a baseline logistic regression
    - Make improvements with each successive version of the model
  - Preprocessing
    - Set missing categorical values to False
    - Algorithmically filter out highly correlated sets of features
      - Keep the feature with the highest variance from each set
      - Avoids multi-collinearity
    - Perform slight 95% Winsorization to reduce the influence of outliers
    - Apply standard scaling, centering on the mean and scaling to standard deviation
  - Key Parameters
    - Balanced class weights
    - L2 regularization to reduce overfitting

Top 4 Features for Prediction

Ordered by magnitude:

1. n_employed

   Portuguese employment count

2. prev_success

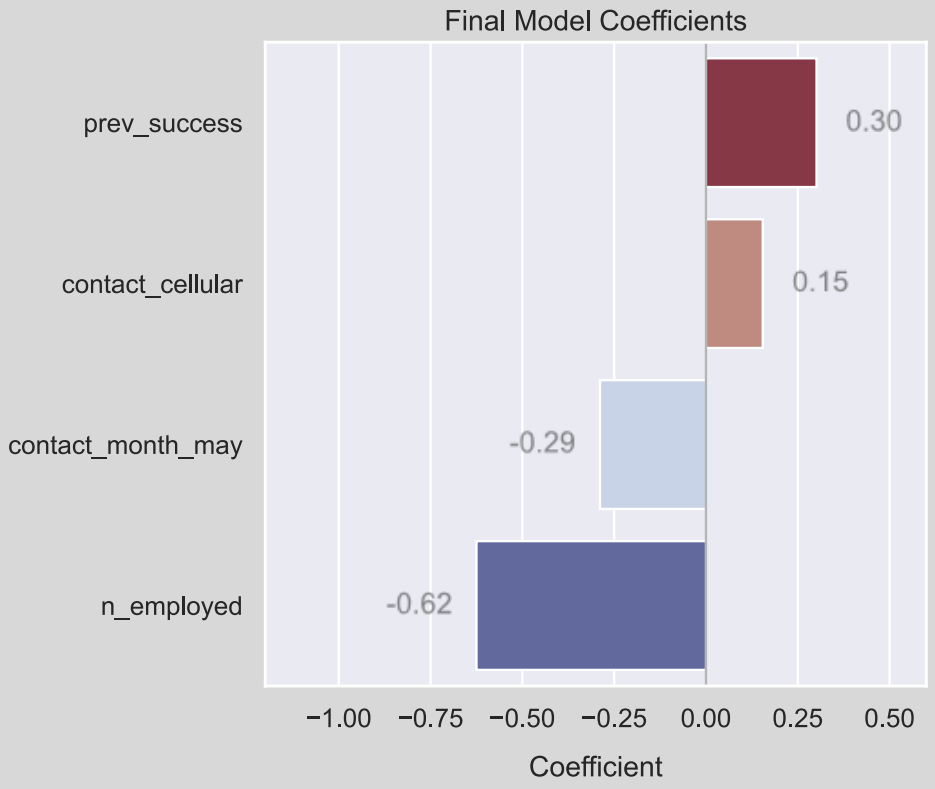
   Customer previously opened an investment account as a result of marketing

3. contact_month_may

   Contacted in May

4. contact_cellular

   Contacted on a cell phone as opposed to a landline



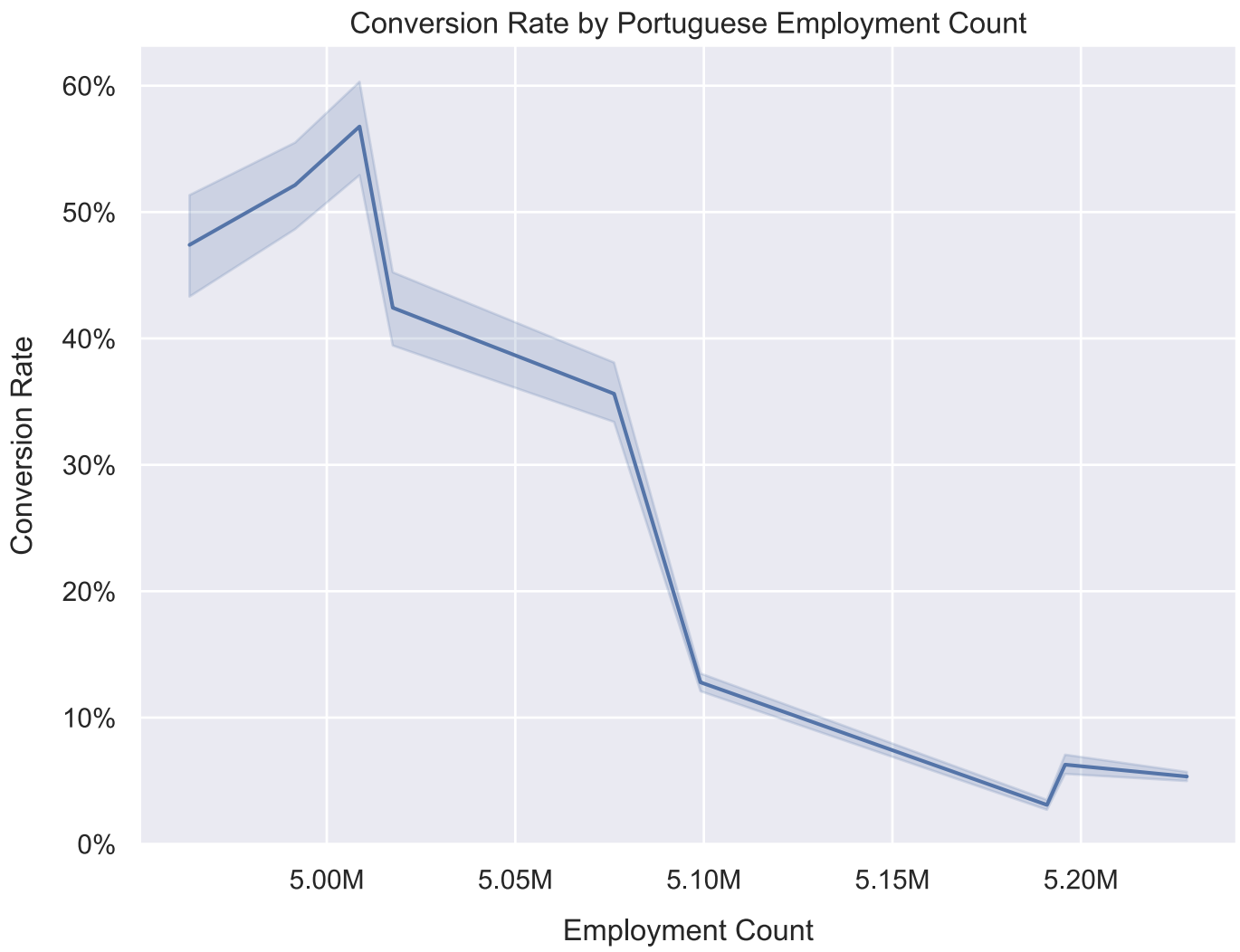
Push Hard When Employment is Low

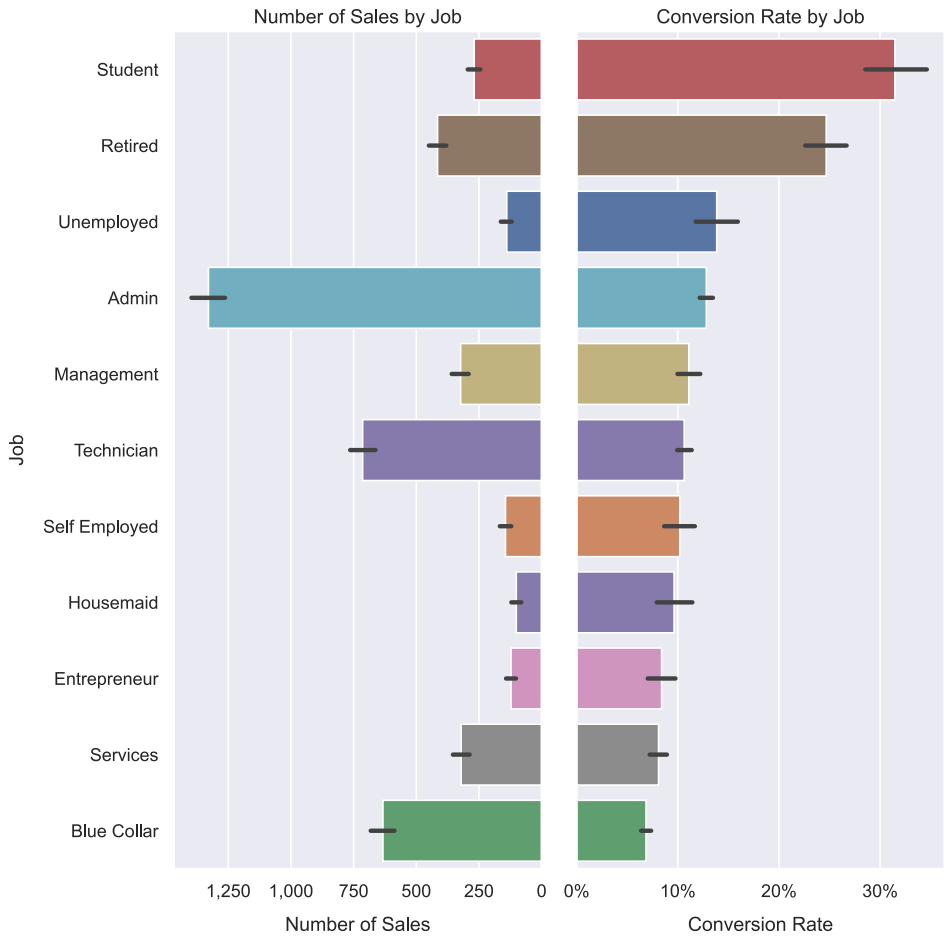
- Undoubtedly, a very strong relationship
- The underlying mechanism behind this relationship is unknown to me
- A surprising discovery resulting from my model
- What is the conversion rate for unemployed people?

Recommendation:

Pump resources into marketing when employment is **low.**

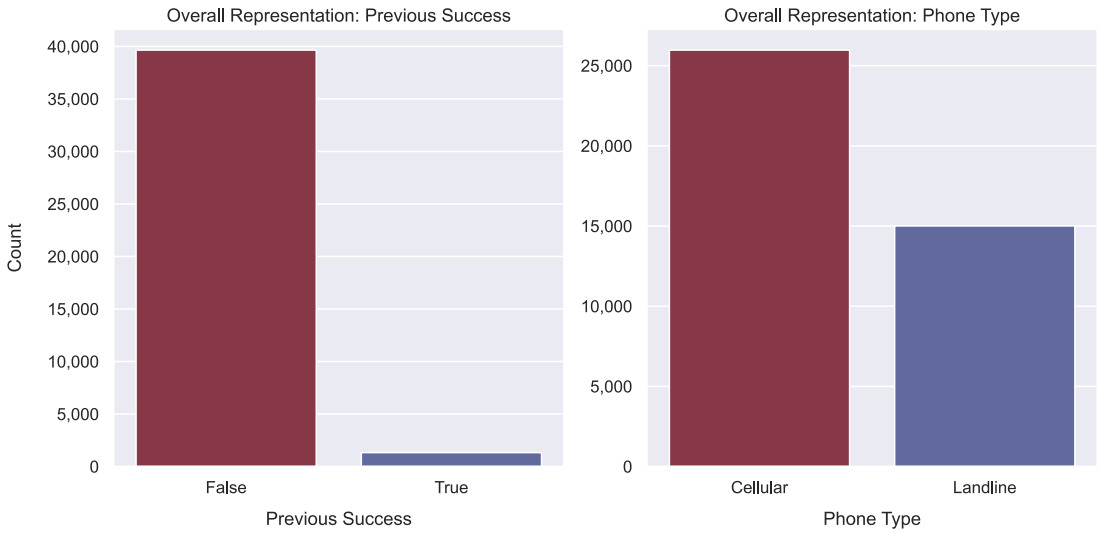
Relax your efforts when employment is **high**.

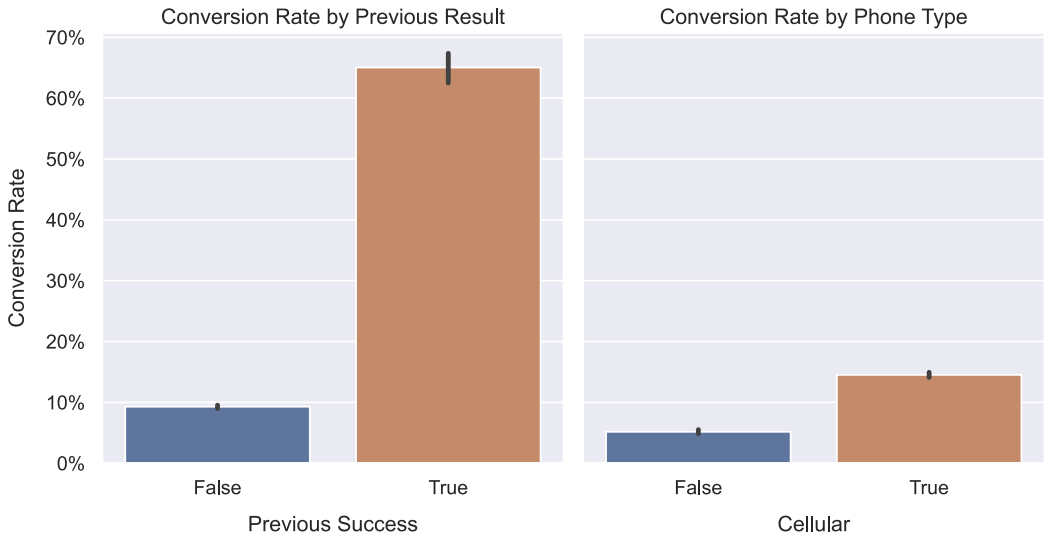




Unemployed People Don't Invest

- The conversion rate for the unemployed is surprisingly high, but not that high
- The total sales to unemployed people is predictably low
- Students and retirees have the highest average conversion rate
- Administrators and technicians have the highest total sales



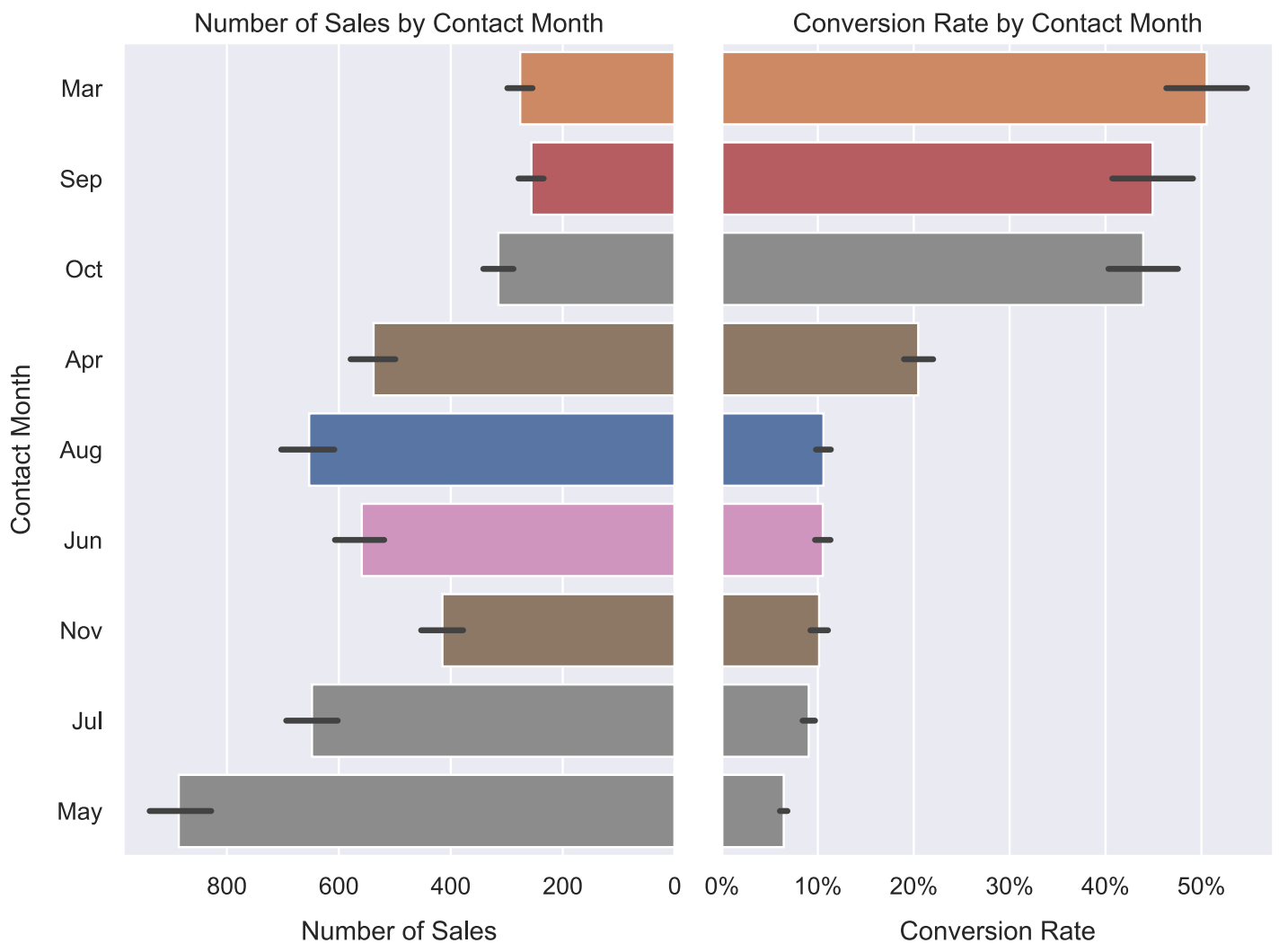


Previous Success is a Major Predictor

- Previous success is one of the biggest predictors of future success
- Unfortunately, only about ~14% of the customers were previously contacted, and of those only ~3% made a deal
- Cell phones are more popular than landlines, but other than that it's unclear why they have a relationship with conversion

Recommendation:

Spend your energy and resources returning to previous customers who were receptive to marketing.

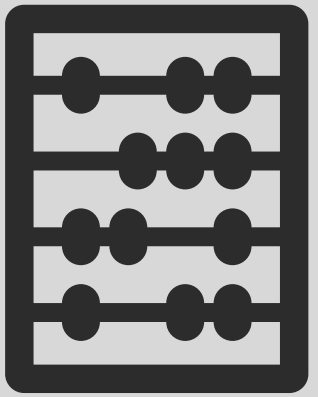


Relax in the Summer

- May has the highest total sales by the lowest average conversion rate
- It also has by far the most data points of any month, at nearly 14,000
- The summer months in general have the highest representation in the dataset

Recommendation: don't put as much energy into Summer marketing as you're used to.

Future Work: More Models



Build and optimize models of different types:

1. Random Forest Classifier
   - No multi-collinearity issue
   - Many hyperparameters to tune
2. Linear Support Vector Classification
   - Can accommodate datasets of this size, unlike other SVMs
3. K Neighbors Classification
   - No multi-collinearity issue
   - Few hyperparameters to tune, but less sophistication

Experiment with feature engineering and joining with other datasets.

- Perhaps other strong positive features like ‘prev_success’ can be engineered
- Consider whether new features will be highly correlated with other features, unless using non-linear model

Thank you!