

Welcome!

CDIPS



Data Science  
Workshop

# Real Estate Marketing for First.io

Alexander Appleton  
Liang Chen  
Tetiana Dadakova

Mentor: Nathan Lande



# Finding a Needle in a Haystack

## Our gigantic dataset

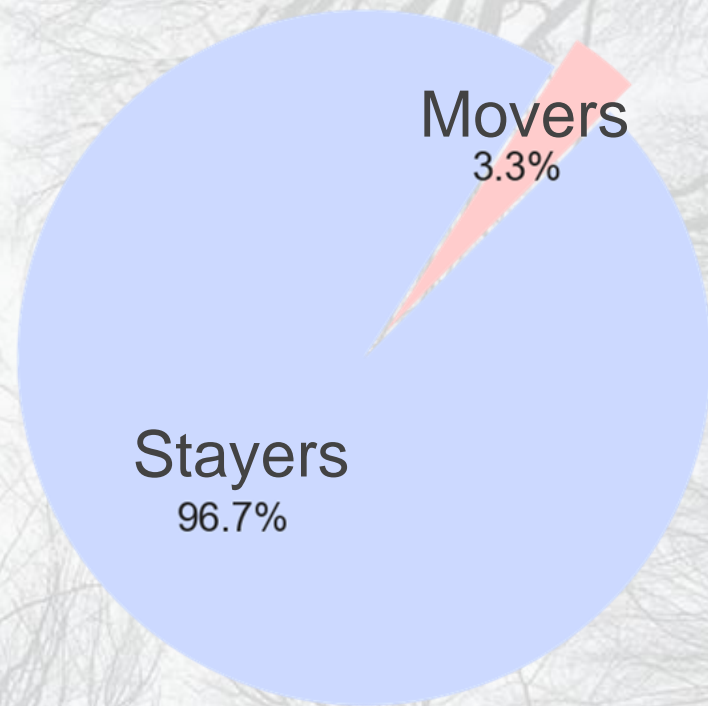
- 700+ features
- ~ 2,000,000 households in the state of NC
- Two time records: 09/2016 and 06/2017

## Our goal

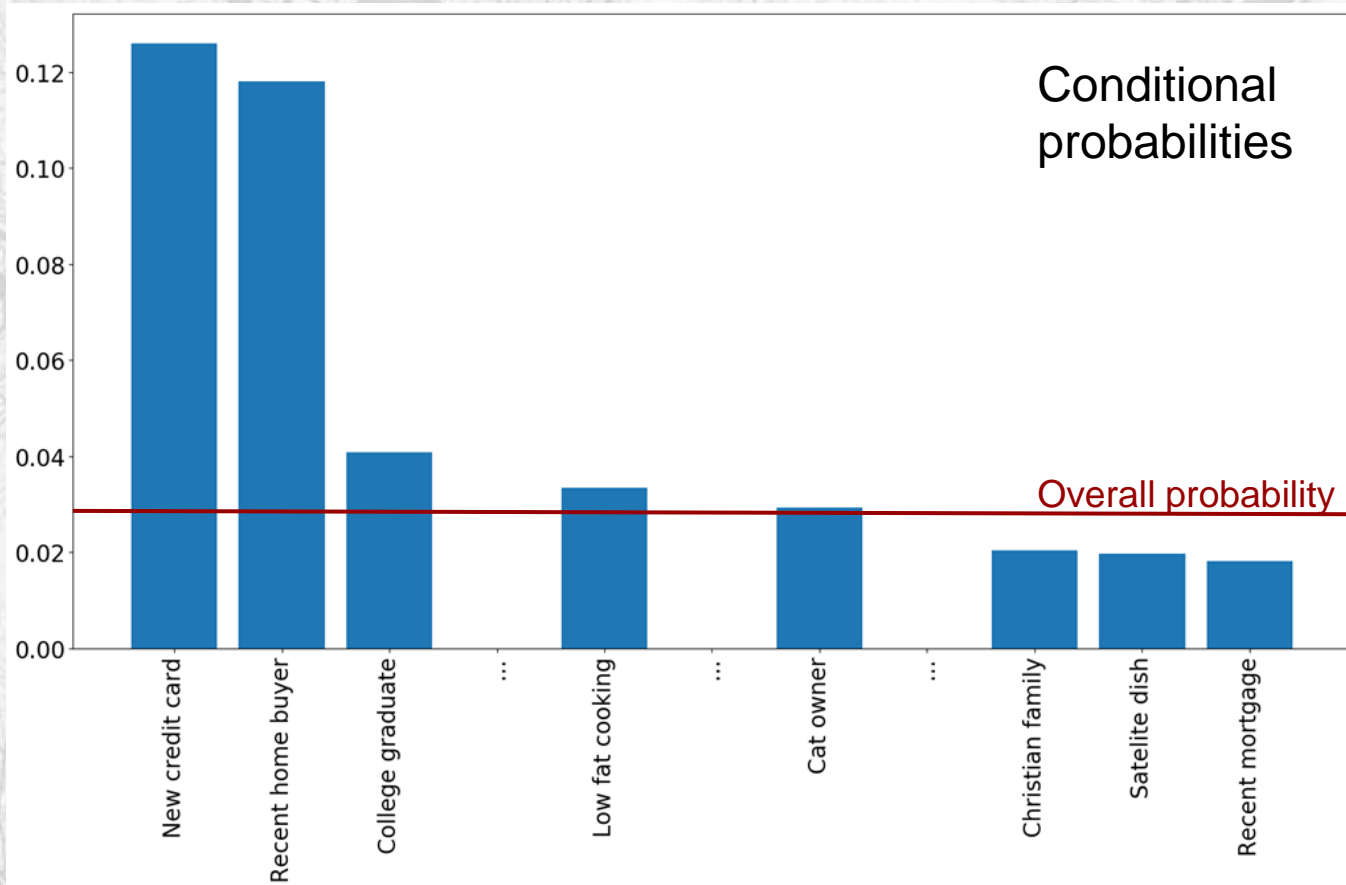
- Predict and validate who has moved between the period of 09/2016 and 06/2017

## Computational tools (BIG challenge)

- Cloud computing resources:
- Distributed file system:
- Distributed computing environment:

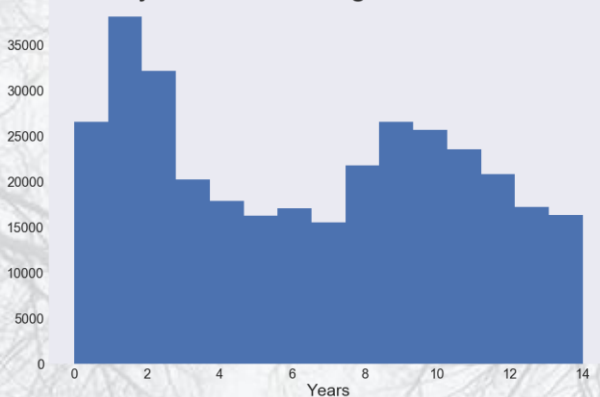


# Explorative Analysis - Binary Features

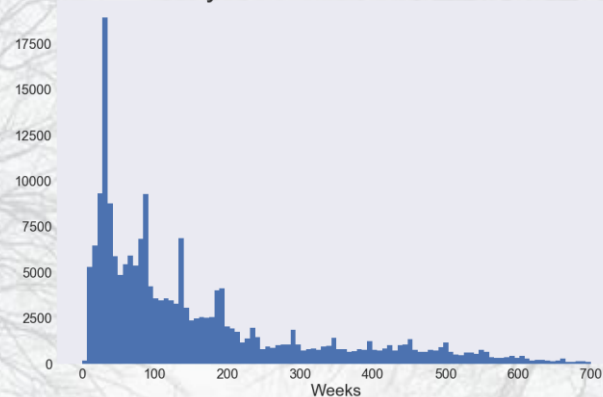


# Explorative Analysis - Numerical Features

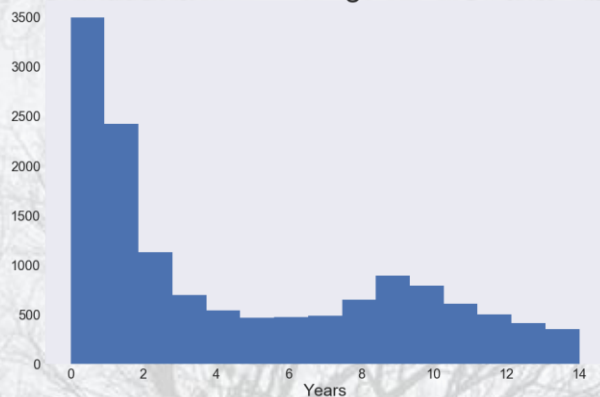
Stayers - Home length of residence



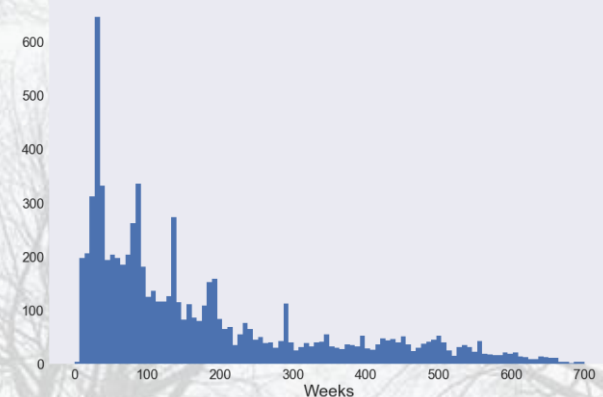
Stayers - Last online order



Movers - Home length of residence



Movers - Last online order





# Hyperparameter Tuning - Logistic Regression

## Hyperparameters

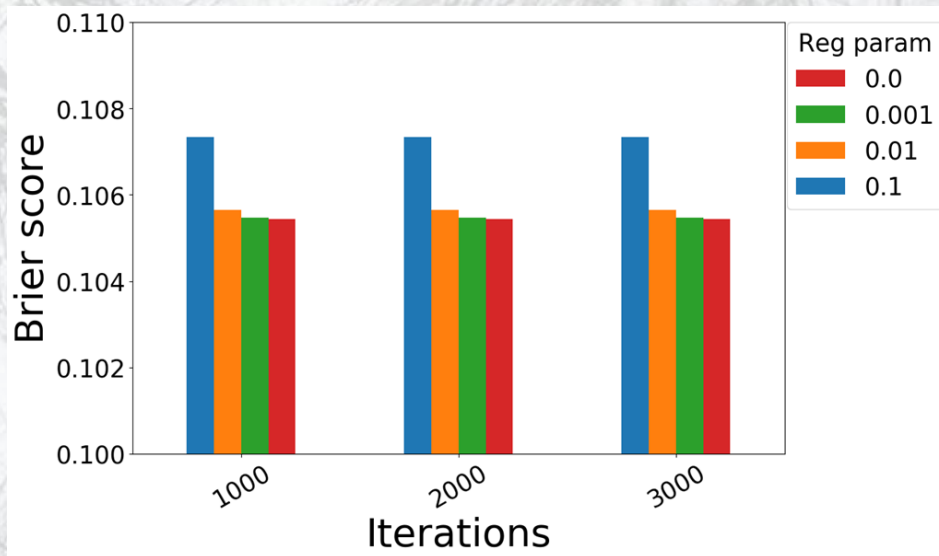
1. Regularization coefficient ( $\lambda$ )
2. L1/L2 regularization ratio ( $\alpha$ )
3. # of iterations in gradient descent

## Lessons Learned

1. No regularization is the best
2. LR model underfits the data
3. Need non-linear models for better performance

## Model metric

$$\text{Brier score} = (\text{probability} - \text{label})^2$$



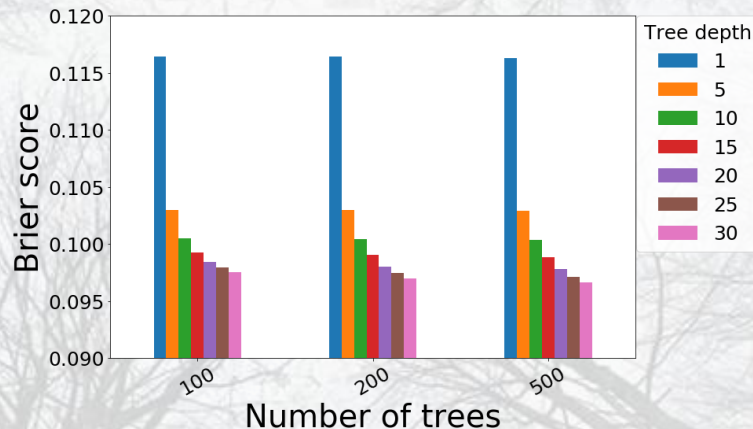
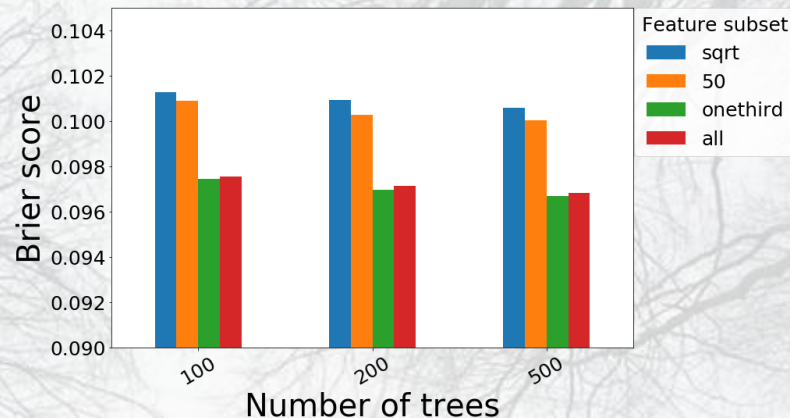
# Hyperparameter Tuning - Random Forest

## Hyperparameters

1. # of trees in the forest
2. Max depth of trees
3. # of features

## Lessons Learned

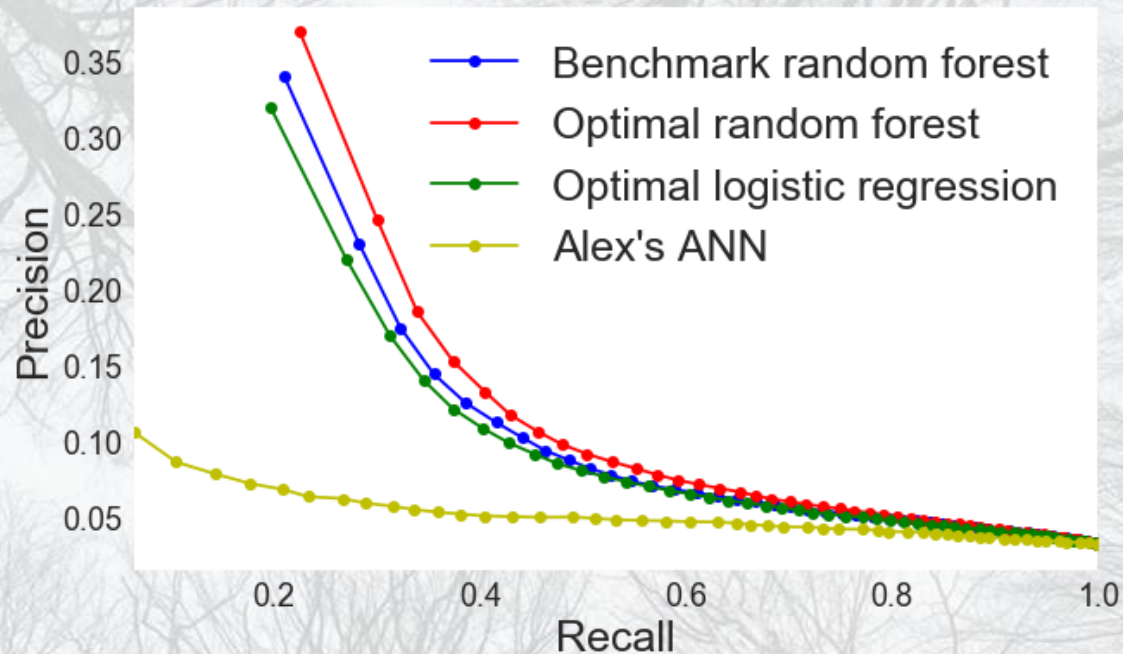
1. Increasing tree depth and using  $\frac{1}{3}$  of total features are 2 critical factors in improving model performance
2. Increasing # of trees has no effect



# Measuring Model Performance - P/R curve

$$\text{Precision} = \frac{\text{True positives}}{\text{All positive predictions (model accuracy)}}$$

$$\text{Recall} = \frac{\text{True positives}}{\text{Total positive targets (model sensitivity)}}$$





# Summary and Path Forward

## Conclusion

- Our optimized RF model achieves 7% improvement in Brier score over the production model

## Immediate Impact

- Insights already applied to the national model

## Long Term Prospects

- Moving decision-making is inherently a time dependent process. A time series model may offer unique perspectives and advantages

