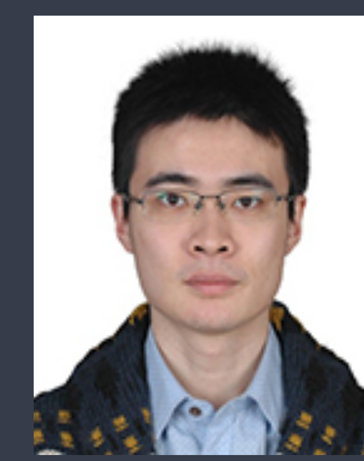
Machine Learning Methods in Credit Risk

MIT MANAGEMENT BUSINESS ANALYTICS

McKinsey&Company

2018 Capstone Project (Boston)



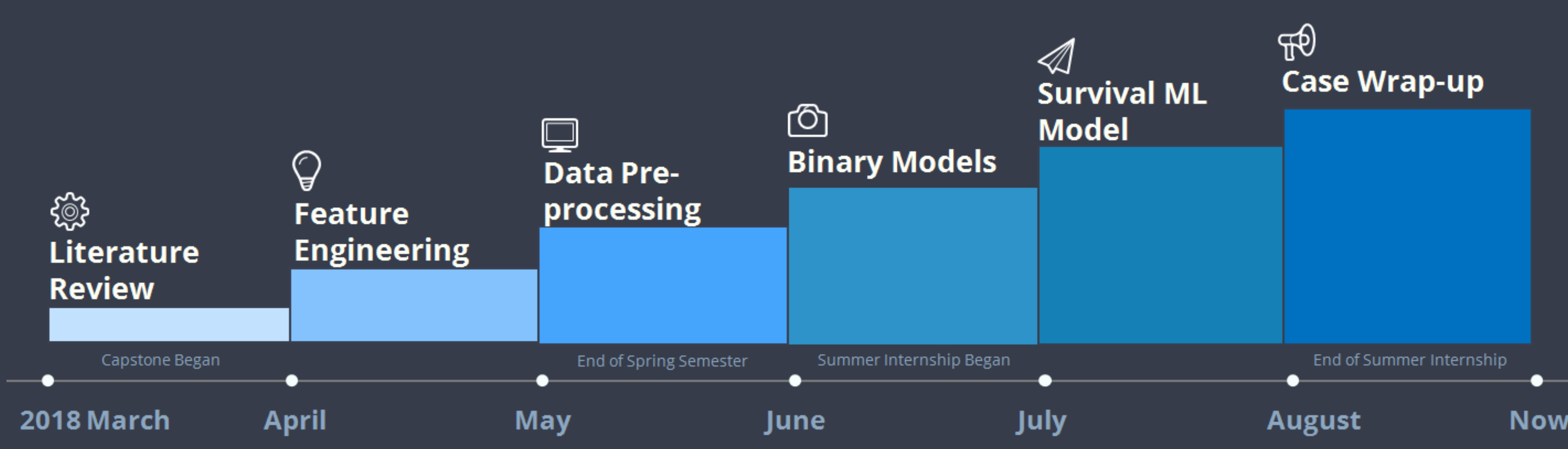
Scott Wang
MBAn
hswang@mit.edu

Tim Yang
MBAn
timsyang@mit.edu

Colin Fogarty
Faculty Advisor
cfogarty@mit.edu

Sean Lu
PhD Mentor
haihao@mit.edu

Capstone Timeline



Problem Statement

The main interest was to help bank determine whether to grant loan depending on the risk of the mortgage. Our goal was to develop a robust model to predict default using available data at the time of the house mortgage application.

Definition of Default

Data Sources

Fannie Mae

Mortgage Loan with terms 10-30 years, acquired by Fannie Mae **2004-2013**

Housing Price Index

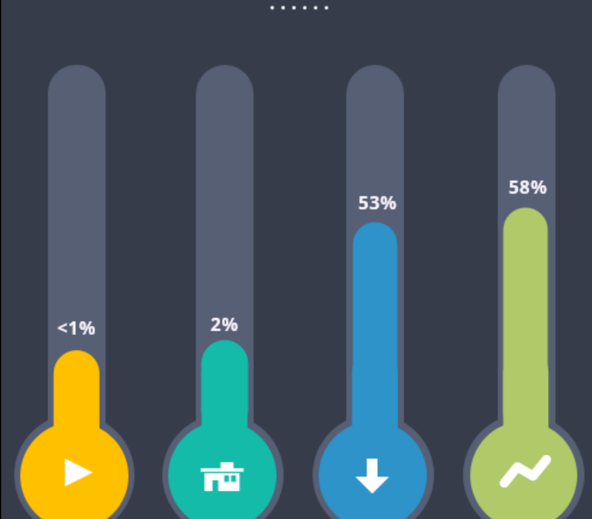
Financial Market (stock, bond, gold price)

Labor Market Data

Methodology, Data Processing And Performance

Sensitivity

In Penalized LR



Baseline

Penalized LR (Elastic net)

Added Feature (HPI)

House Price Index

Downsample

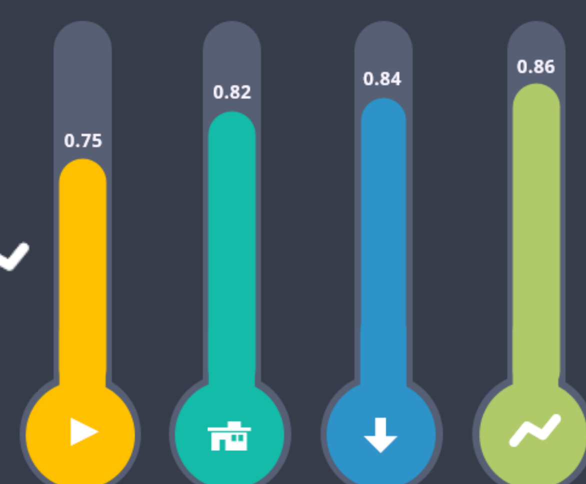
House Price Index

Added more features & data

Labor Statistics, Macroeconomic Variables (Stock, Bond, Gold Price)

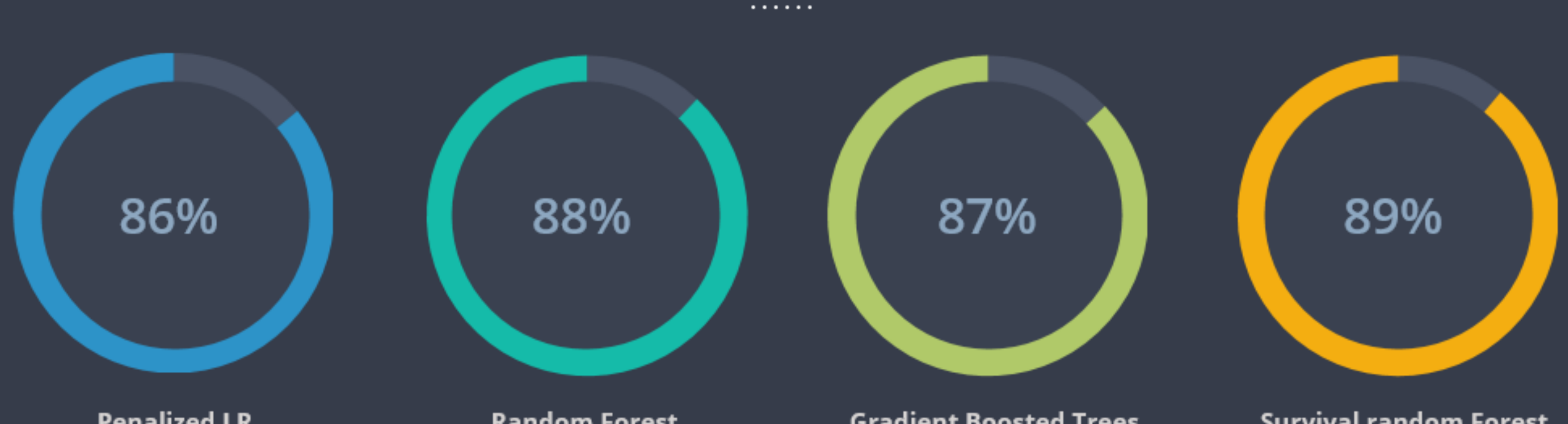
Out-of-Sample AUC

In Penalized LR



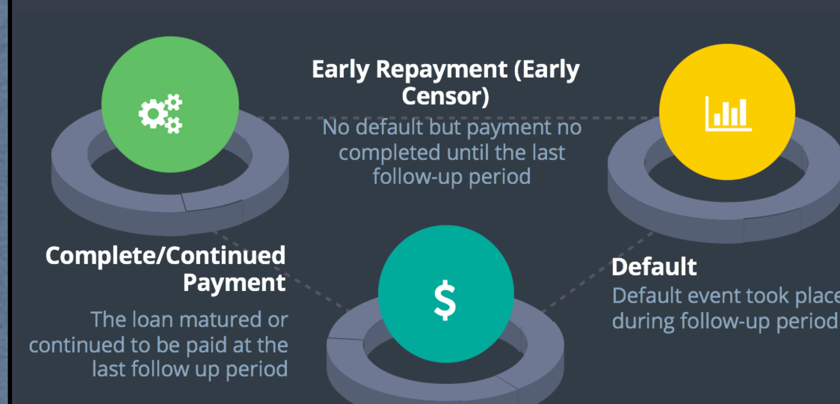
Algorithm Performance

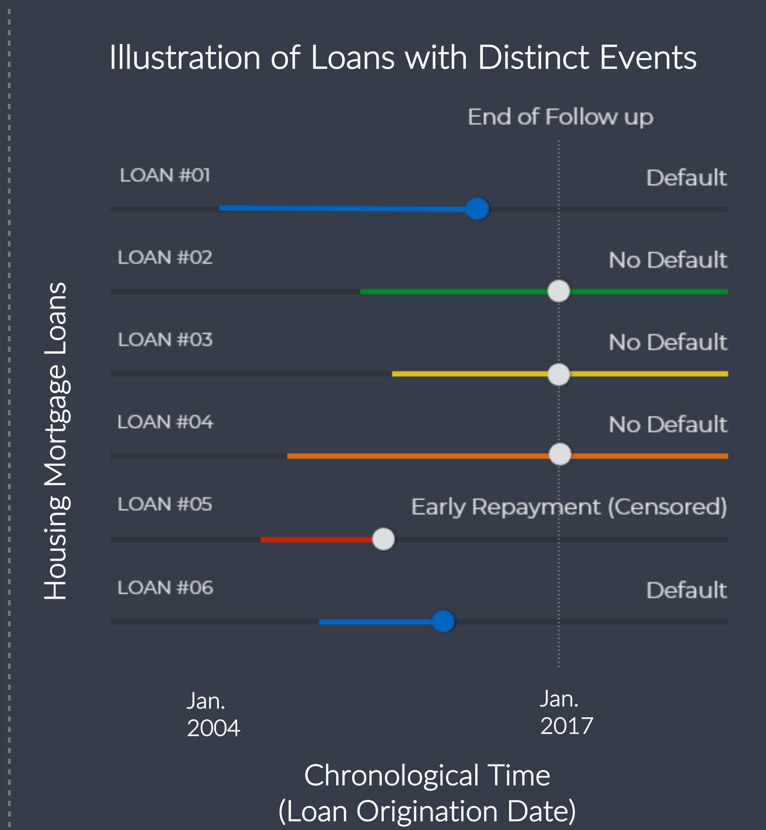
In AUC



Survival Analysis in Default Prediction

- **Classification** (Predict the state): Event/no event/censored
- **Regression** (Predict time to default): Time to event/no event/censor





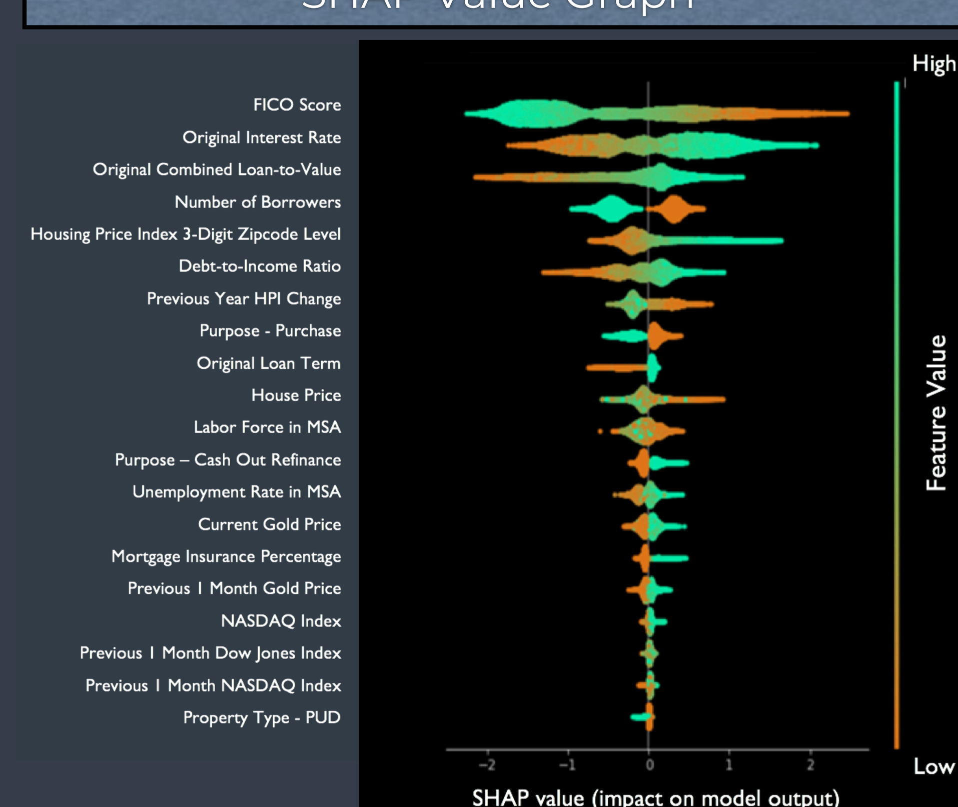
Interpret Machine Learning Models

	Method	Description
Classical Feature Importance Methods	Weight (Split Count)	Calculated as the total decrease in node impurity (weighted by the probability of reaching that node) averaged over all trees of the ensemble.
	Gain (Mean Decrease Gini)	Calculated as the total decrease in node impurity (weighted by the probability of reaching that node) averaged over all trees of the ensemble.
	Permutation	Calculated as the decrease of accuracy of the model predicted on intact OOB samples and OOB samples where the values of a specific variable are randomly permuted.
	Cover	Calculated as the average coverage (the number of samples affected by the split) of the feature when it is used in trees.
Newer Feature Importance Methods	LIME (Local Interpretable Model-Agnostic Explanations)	Globally breaks down the feature components and fits in local (single data-point) level to visualize the effect and direction for each variable on this single data point.
	Tree SHAP (Shapley Additive Explanation)	Ranks feature importance by calculating mean absolute value of SHAP values, which average differences in predictions over all possible orderings of features for each individual observation.

Classic Feature Importance

Cons:
- No magnitude/direction
- Not consistent
- Hard to validate with prior knowledge

SHAP Value Graph



Color → **Magnitude**

SHAP Value → **Direction**

Thick → **#Stacked Individuals**

Pros:
- Has magnitude/direction
- Accurate and consistent
- Easier to interpret complex relationships

Business Impact

Key Conclusions

Machine learning models especially survival models add value and valuable insights

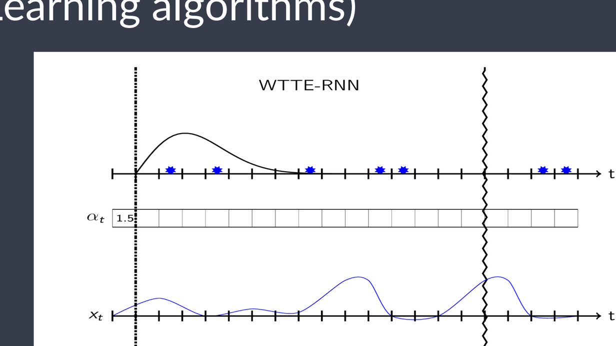
The results of this study will be used to build comprehensive and accurate credit risk models for future customers

Economic Impact via Optimization

Baseline	Machine Learning	Survival Machine Learning
Ground-Truth Economic Loss	Cut 80% Ground-Truth Loss	1% Further Improvement

Future Directions

- Choose cutoffs via portfolio loss optimization
- Real-time portfolio risk monitor with time-varying covariates (through Deep Learning algorithms)



Dynamic Time to Default Estimation