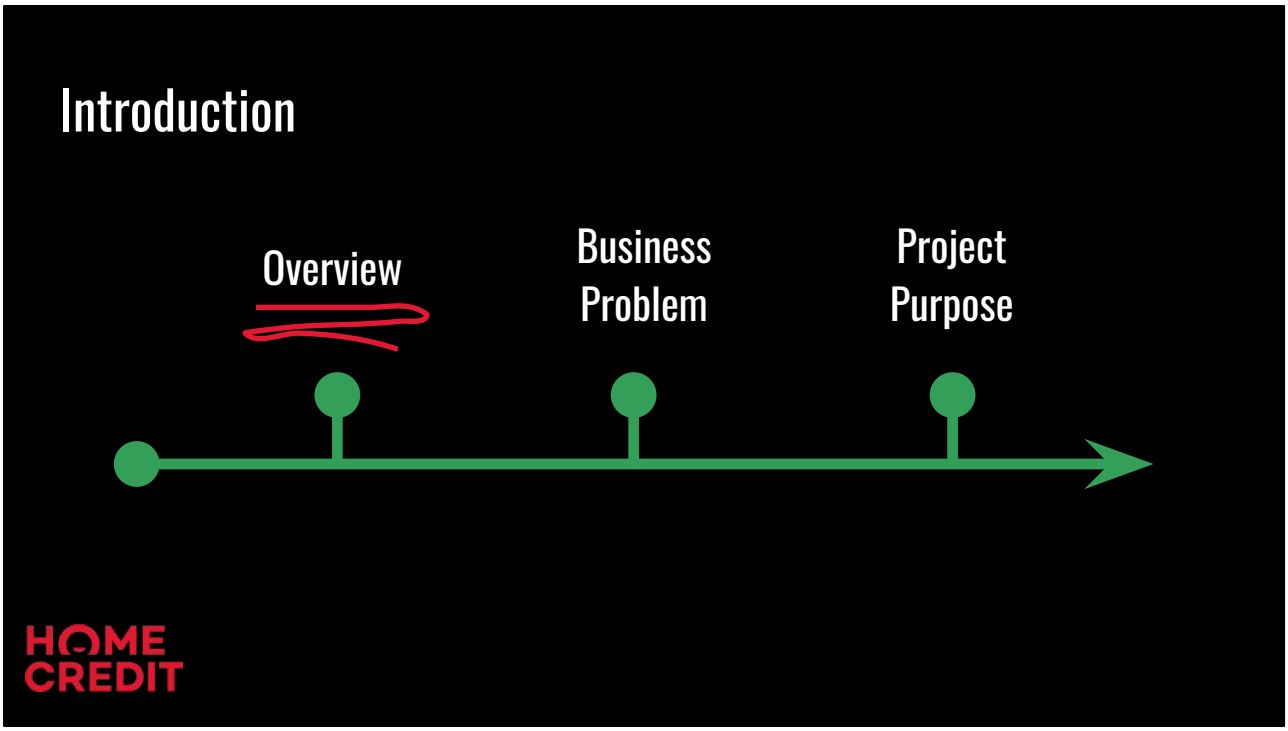
Whitney - 1 min

“Hi everyone, my name is Whitney Holt and these are my teammates: Maddie Lee, Alexia Wells, and Leah Ekblad.

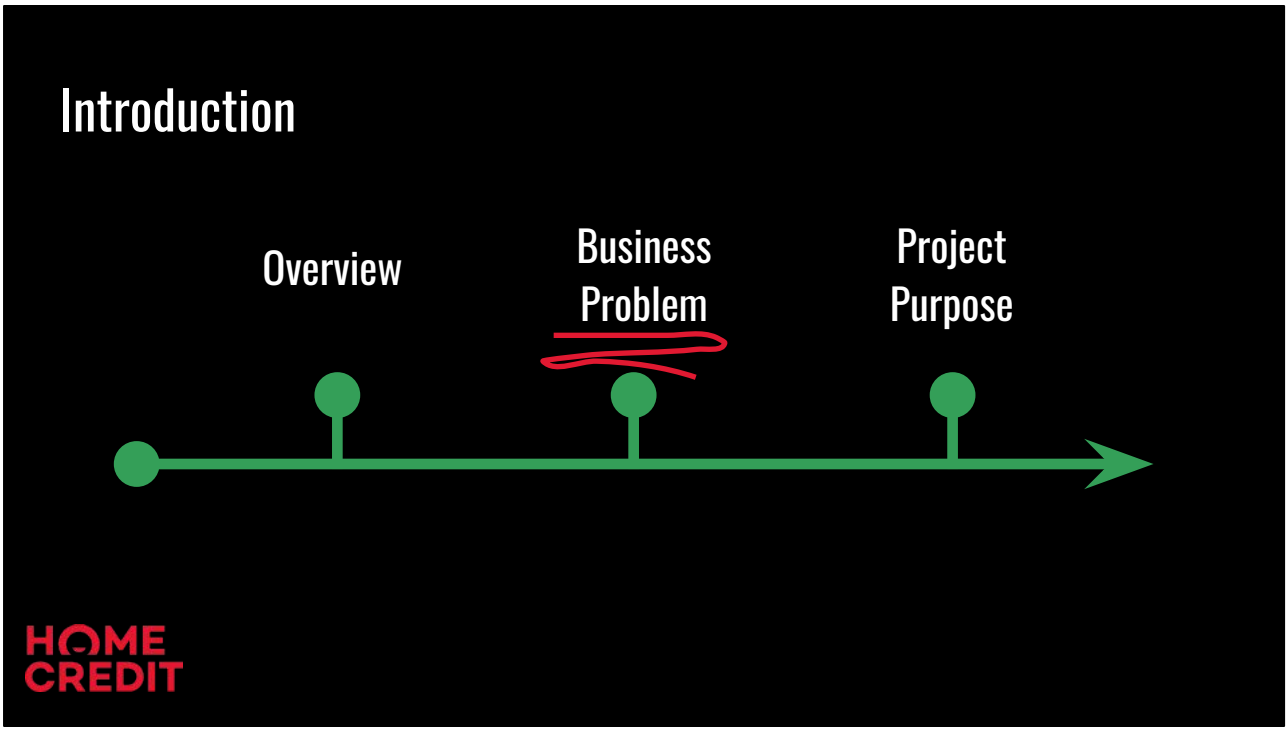
Today, we are excited to speak to you about the possibility of enhancing loan approval processes at ***Home Credit*** *through predictive modeling.”*



Whitney - 1 min

"*Before we jump in, I'd like to provide a overview of our presentation:*

- *This presentation will cover*
  - ***Business Problem***
  - *and subsequent* ***Project Purpose***
  - *a promising* ***Ensemble Model***
  - *related* ***Business Implications***
  - *and some* ***Potential Next Steps***
- *Any remaining time will be used to answer* ***Questions****."*



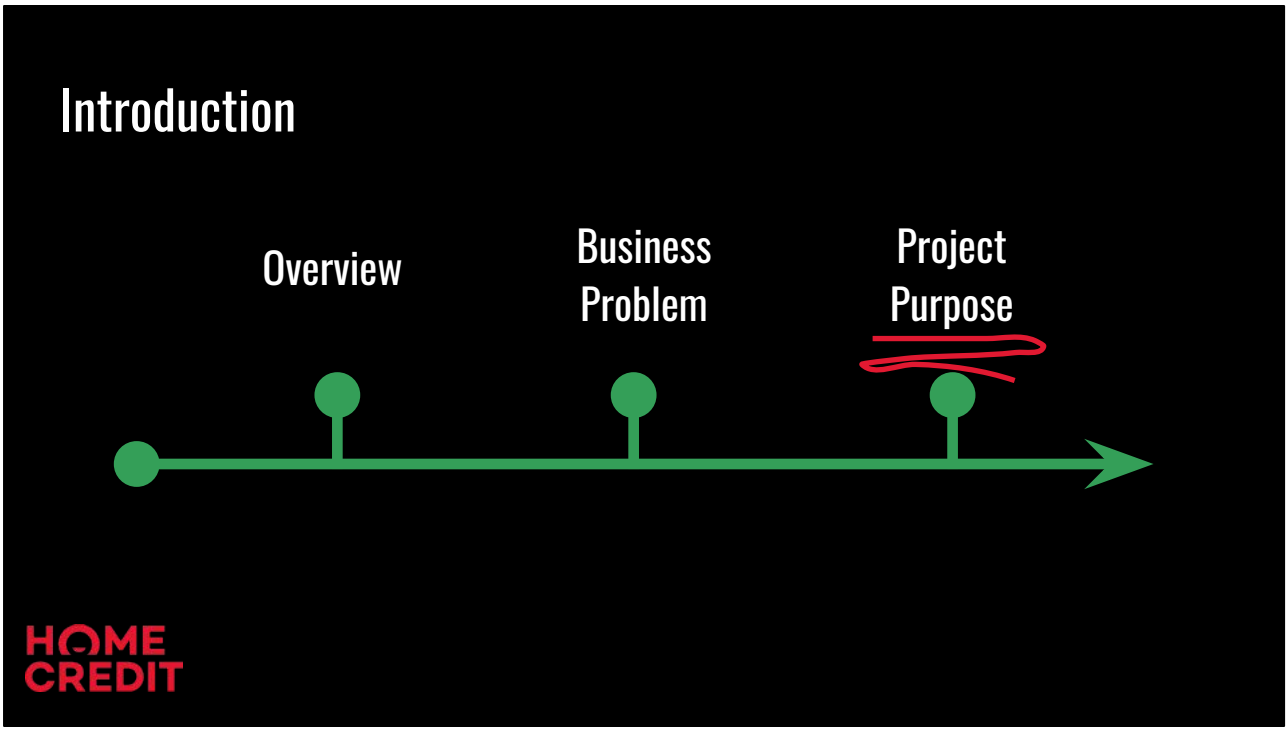
Whitney - 1 min

- ***Business Problem***:

  "Financial institutions and lenders often use a customer's credit history to approve loans and set interest rates.

  Individuals that lack credit are often denied loans or are vulnerable to predatory lenders even if that individual is capable of repaying their loans.

  This creates a loss of opportunity to both the borrower and creditor, which ***Home Credit*** *seeks to overcome."*



Whitney - 1 min

- ***Purpose of this Project:***

  "This project aims to build a predictive model that evaluates loan repayment capability for individuals with limited or no traditional credit history.

  A successful model will allow ***Home Credit*** *to assess applicants based on alternative data, expanding credit access responsibly."*

- *"Now I'll hand it over to* ***Alexia****, who will talk about our team's* ***Ensemble Model****."*

Ensemble Model

- Models
  - Random Forest
  - Bayesian Additive Trees
  - Logistic Regression
  - Extra Trees
- Kaggle AUC: 0.70311
- Runtime: 1.05 hours

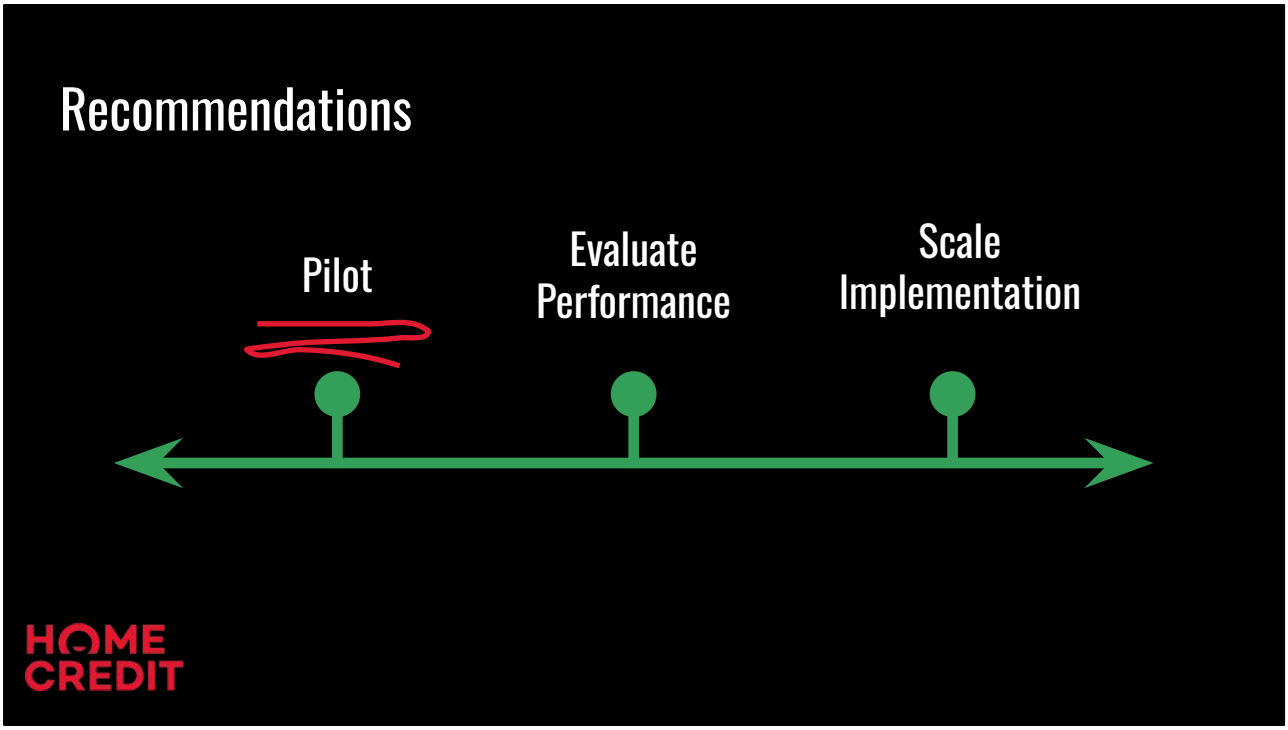
HOME CREDIT

The best performing model was an ensemble method using stacking! This method included extra financial feature engineering. We combined predictions from four different models - random forest, BART, logistic regression and extra trees. This method is likely the highest performing because the individual models themselves were all high scoring on Kaggle between .69-.70. When combined with the meta-learner, which was a lasso penalized regression, the most important features were selected! The Kaggle score was 0.70311 and the runtime took 1.05 hours, using just 5% of the training data. It is possible that using the full dataset would cause for a higher Kaggle score, the downside is that it would certainly increase the model runtime. One way to decrease the runtime, would be to use servers to run the models. The final decision for how to implement everything would be up to the company based on their priorities.

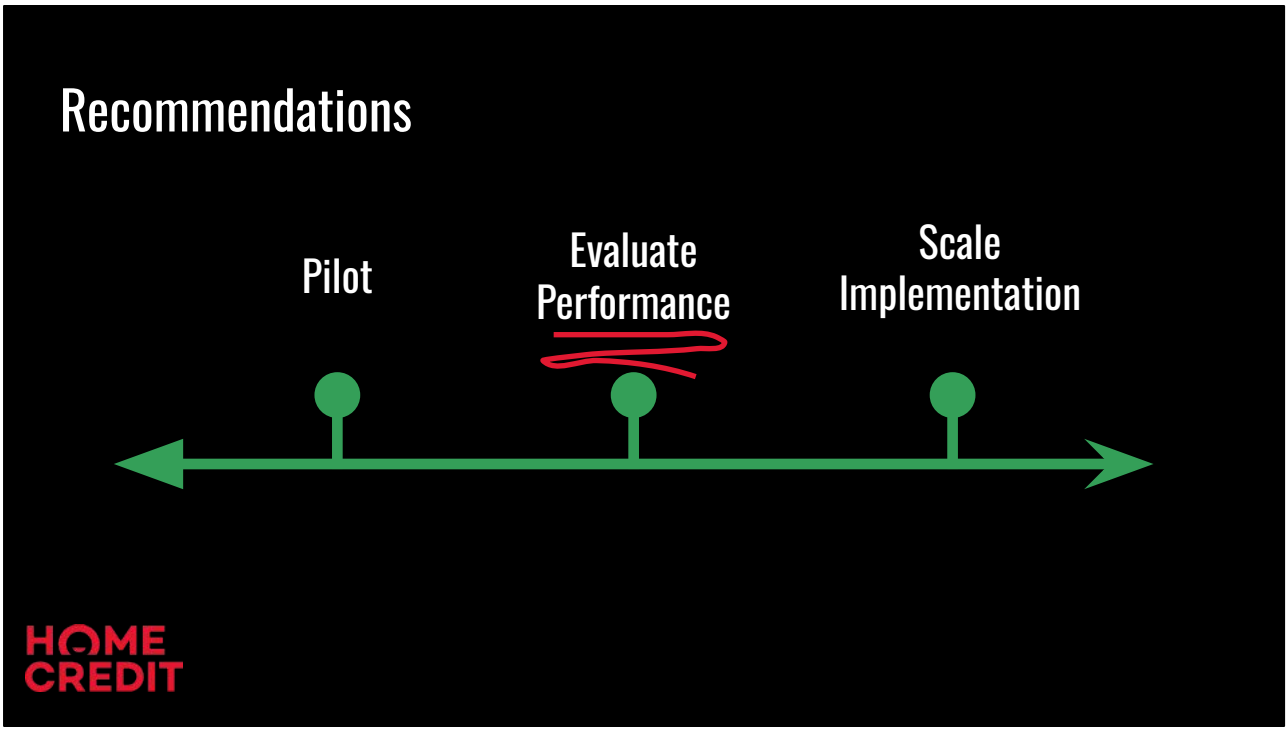
Recommendation

HOME CREDIT

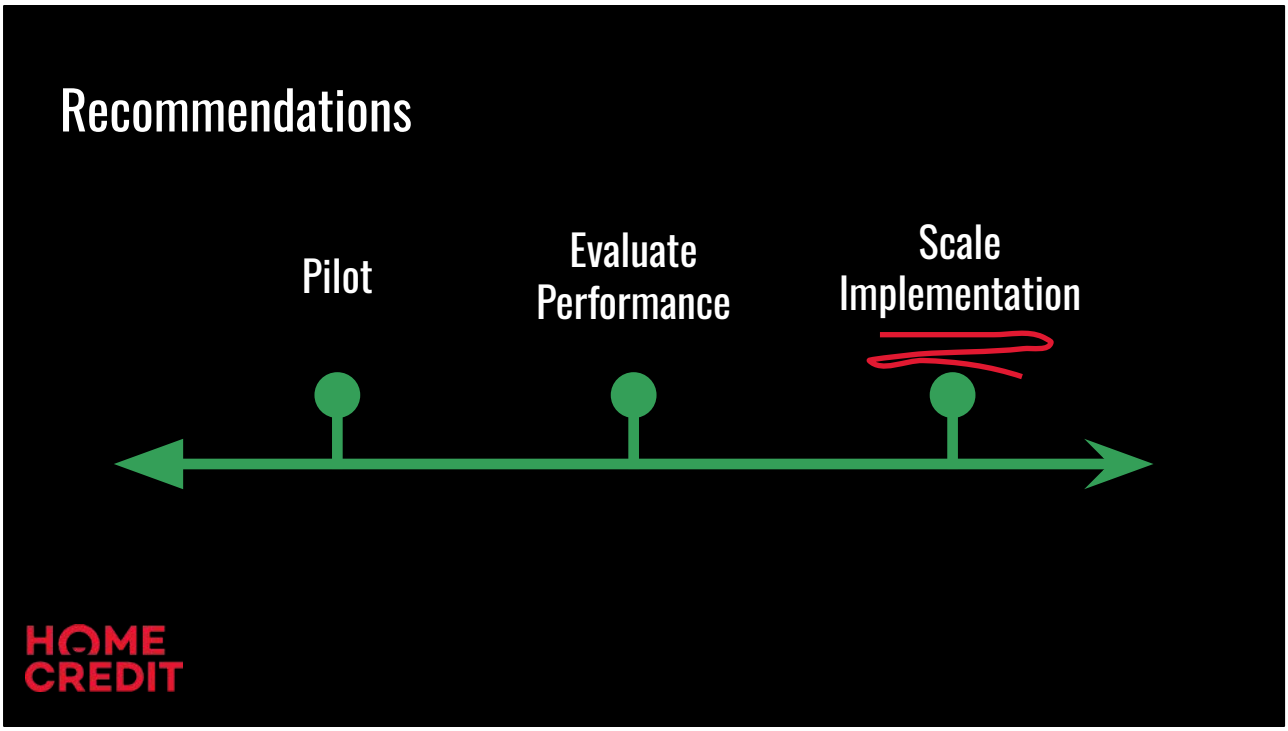
We believe Home Credit should implement the model as a secondary screening tool for applicants lacking traditional credit data



The company should first pilot the model with a subset of applicants lacking credit scores.



Then, evaluate performance against key metrics like approval rates, default rates, and customer feedback.



And finally, scale implementation across all branches based on pilot results.

*Next, **Maddie** will cover Business Implications.*

Business Implications

HOME CREDIT

Thanks Alexia

Now that we have a recommendation on how to implement our model, let’s talk about the possible benefits

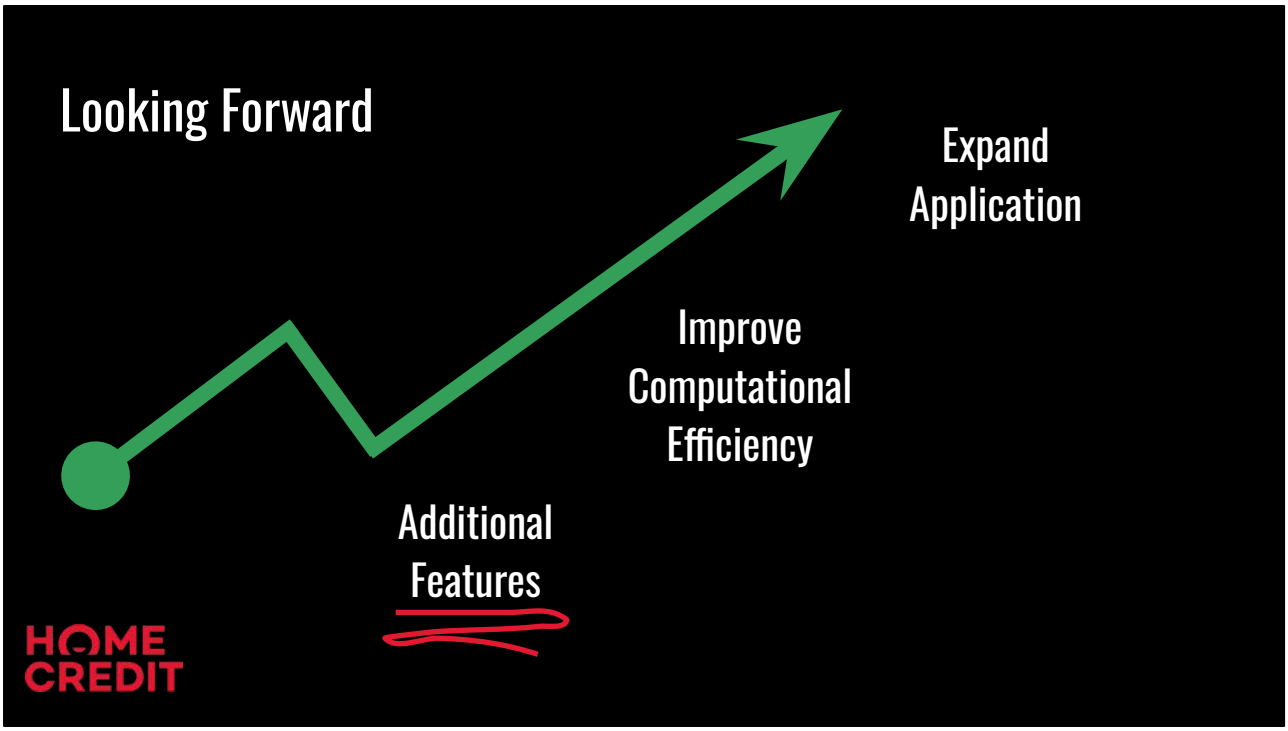
Our model empowers Home Credit to expand its lending reach by identifying and approving loans for customers who lack traditional credit scores. This not only addresses the significant gap in access to credit but also supports financial inclusion, enabling Home Credit to responsibly grow its customer base by serving underserved markets.

With this model, Home Credit can increase loan approvals while maintaining rigorous risk standards. By unlocking access to untapped markets with high-potential borrowers, the company can drive significant revenue growth. This approach ensures that business expansion aligns with sustainable and responsible lending practices.

Maintaining risk control is central to our approach. The model achieves a strong balance between increasing loan approvals and keeping default rates low. By leveraging predictive insights, Home Credit can identify low-risk borrowers, ensuring that growth is both profitable and sustainable

Implementing this model positions Home Credit as a leader in innovative and inclusive lending. It strengthens the company's competitive advantage by aligning with fair lending practices and emphasizing customer-first strategies. This approach enhances brand equity and establishes Home Credit as a socially responsible and forward-thinking lender in the market.

I'll pass it over to **Leah**, who will talk about next steps.



Leah - 1 min

“Thanks Maddie!”

*“Looking forward, let’s discuss some **potential next steps**:*

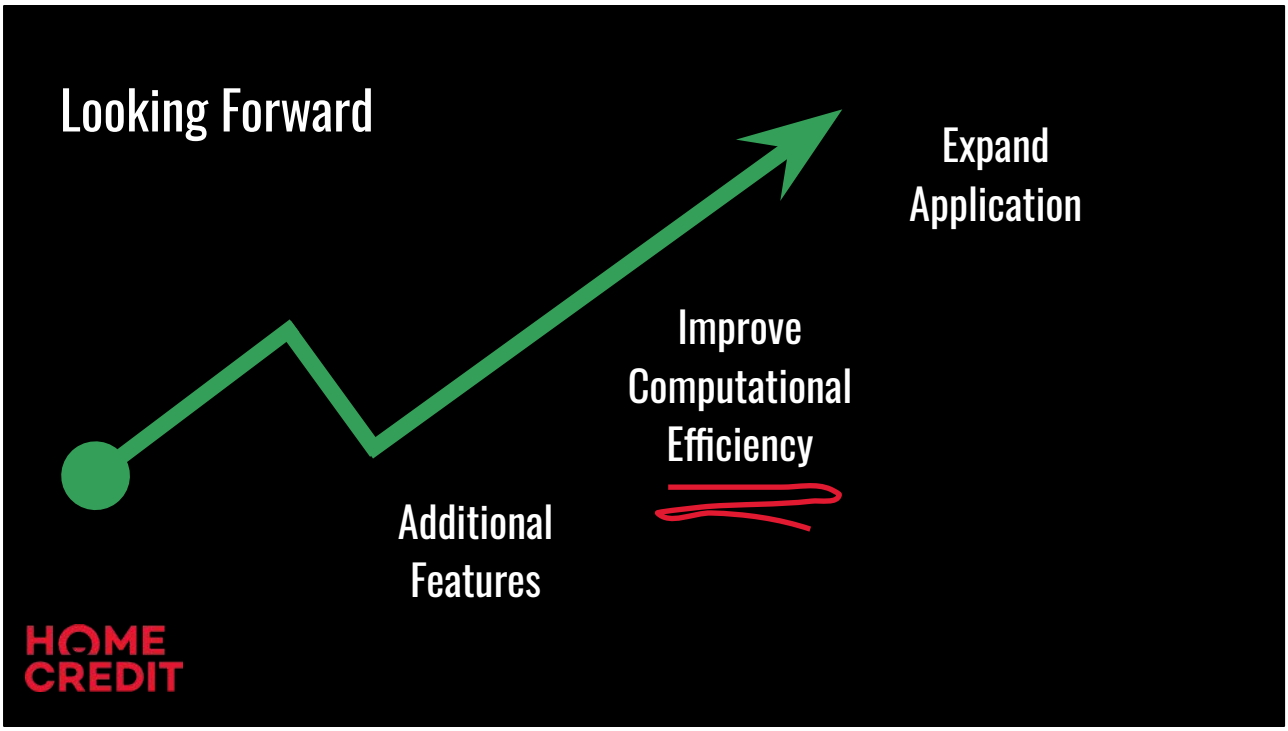
*First, the team could **explore additional features** to include in the model such as **transaction-level data** or **social determinants** to further **refine and improve predictions**.”*

Transaction-Level Data:

- **Transaction Amounts**: The frequency and size of transactions (e.g., daily spending habits, large one-time purchases).
- **Transaction Categories**: Types of purchases (e.g., groceries, entertainment, utilities), which could indicate financial behavior or lifestyle.
- **Payment Methods**: Whether transactions are made using credit cards, debit cards, or digital wallets, which may reflect financial habits.
- **Merchant Information**: Where transactions occur (e.g., high-end retailers vs. discount stores) could provide insights into a person's financial profile.
- **Transaction Recency/Frequency**: How recently and how often a customer makes transactions, which could indicate financial stability or risk.
- **Transfer or Payment Patterns**: Frequency of sending money to other accounts (e.g., remittances, peer-to-peer payments).
- **Debt Repayments**: Regular payments on loans or credit card bills, indicating financial responsibility or strain.

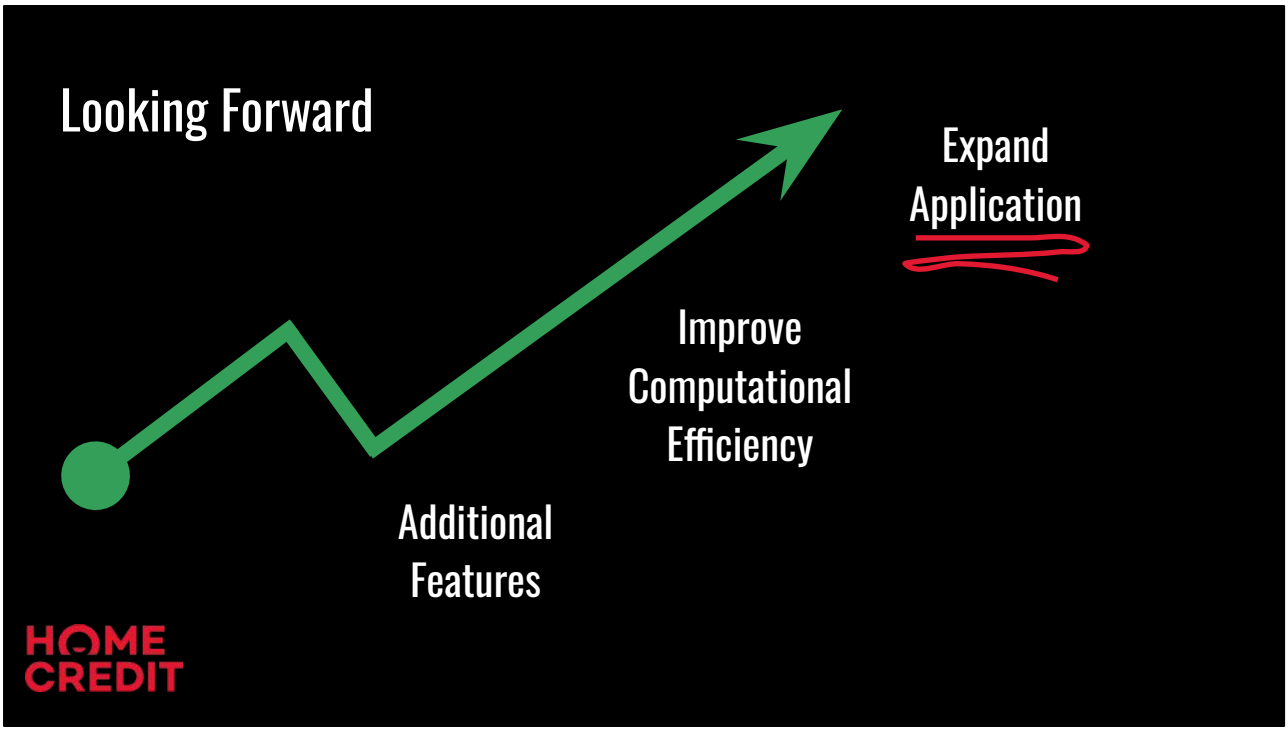
Social Determinants:

- **Income Level**: Household or individual income, a key factor in financial decision-making and risk assessment.
- **Education Level**: Highest degree attained, which may correlate with financial literacy or earning potential.
- **Employment Status**: Employment type (e.g., full-time, part-time, self-employed) and job stability, which influence financial health.
- **Neighborhood/Location**: Geographic location or neighborhood, which can impact access to resources, job opportunities, and economic stability.
- **Family Structure**: Household size, number of dependents, and marital status, which can affect financial priorities and risks.
- **Health and Disability Status**: Health conditions or disabilities that can influence an individual's ability to work or manage finances.
- **Social Support Networks**: Availability of financial or social support from family, friends, or community organizations.
- **Housing Stability**: Whether the individual rents or owns, or if they've experienced housing insecurity, which could correlate with financial stability.
-



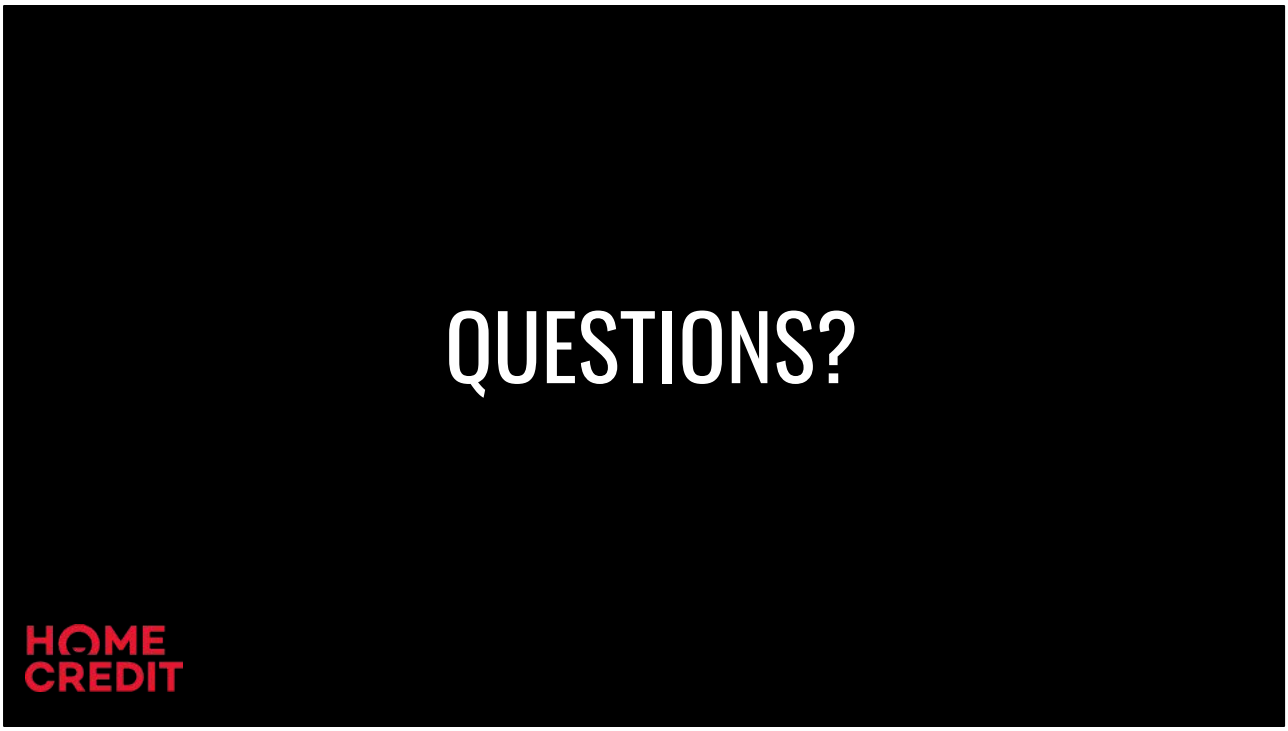
Leah- 1 min

*"Second, the team could **improve computational efficiency** by **reducing processing time**, ultimately **enabling the possibility of making credit decisions in real-time**."*



Leah - 1 min

"Lastly, ***Home Credit*** *may consider* ***expanding the application of this model to other financial products****, such as microloans or insurance, further* ***unlocking even more business potential****."*



Q&A - 1 min

Thank you all for joining, we will now open it up for any questions that you may have

Appendix

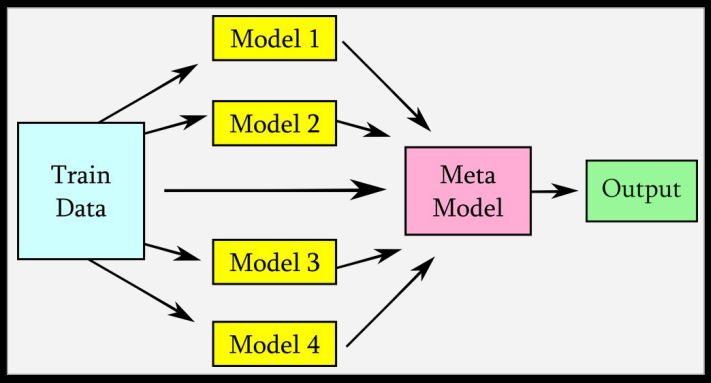
HOME CREDIT

Model Performance Comparison

Model Details		Performance Metrics		
Model Type	**Model Description**	**In-Sample AUC**	**Out-of-Sample AUC**	**Kaggle Score**
Majority Class Classifier	Baseline Model	0.500	0.50	NA
Logistic Regression	Feature engineering, all predictors	0.770	0.66	NA
Logistic Regression	All predictors	0.680	0.65	NA
Logistic Regression	All predictors, stepwise	0.740	0.67	0.677230
Logistic Regression	Feature engineering, all predictors	0.691	NA	0.678960
Logistic Regression	Feature engineering, attribute selection	NA	NA	0.639271
Naive Bayes	Feature engineering	0.680	0.66	0.582870
LASSO Regression	Feature engineering	0.720	0.68	0.676590
Random Forest	500 trees	0.720	0.68	NA
Random Forest	Feature engineering, 500 trees	NA	NA	0.673620
Random Forest	Feature engineering, 100 trees	NA	NA	0.663660
Random Forest	Top 10 features, 500 trees	NA	NA	0.663660
Random Forest	Using LASSO-selected predictors, 500 trees	0.710	0.68	0.676240
Random Forest	MTRY Adjustment, 500 trees	NA	NA	0.673840
Random Forest	Feature engineering, all predictors, auto hyperparameter tuning	0.690	NA	0.698770
Extra Trees	Feature engineering, all predictors, 1000 trees, auto hyperparameter tuning for min_n()	0.690	NA	0.693390
LightGBM Boosted Trees	Feature engineering, all predictors, auto hyperparameter tuning	0.690	NA	0.686760
Bayesian Additive Regression Trees (BART)	Feature engineering, all predictors, auto hyperparameter tuning	0.693	NA	0.701130
Ensemble Model	LightGBM + BART + Logistic Regression	NA	NA	0.691690
Ensemble Model	Random Forest, Logistic Regression, and BART	NA	NA	0.700500
Ensemble Model	LightGBM + Random Forest + Extra Trees + Logistic Regression	NA	NA	0.697650
Ensemble Model	Random Forest + BART + Logistic Regression + Extra Trees	NA	NA	0.703110

Appendix - Stacking Model Predictions

- Goal - Reduce bias and improve predictions
  - Combines predictions of diverse models
  - Meta-learner that makes the final predictions



HOME CREDIT

The best performing model was an ensemble method using stacking! We tried this out because most of the top leaderboard submissions used this method. Essentially, you use several different models to train the data, put all of them into a meta model. Once that is done, you get the finalized predictions.

Appendix

```
# Stacked predictions/ensemble method

# Best performing
my_stack <- stacks() %>%
  add_candidates(rf_CV_results) %>%
  add_candidates(bart_CV_results) %>%
  add_candidates(log_CV_results) %>%
  add_candidates(extra_trees_CV_results)

## Fit the stacked model
stack_mod <- my_stack %>%
  blend_predictions() %>% # LASSO penalized regression meta-learner
  fit_members() ## Fit the members to the dataset

## Use the stacked data to get a prediction
stacked_predictions <- stack_mod %>%
  predict(new_data=test, type = "prob")
```

HOME CREDIT

This is a quick look at what our code looked like. As you can see, you add all the models, and put it into a LASSO penalized regression metal-learner, which performs variable selection and reduces model complexity.

Appendix - Ensemble Model

- Models Used
  - Random Forest - Bagging/boost strapping aggregating
  - Bayesian Additive Trees - Essentially looped boosting algorithm
  - Logistic Regression
  - Extra Trees - Bagging
- Kaggle AUC: 0.70311
- Runtime: 1.05 hours

HOME CREDIT

The best performing model was an ensemble method using stacking! This method included extra financial feature engineering from previous_application.csv. We combined predictions from four different models - random forest, BART, logistic regression and extra trees. This method is likely the highest performing because the individual models themselves were all high scoring on Kaggle between .699-.701. When combined with the meta-learner, which was a lasso penalized regression, the most important features were selected! The Kaggle score was 0.70311 and the runtime took 1.05 hours, using just 10% of the training data. It is possible that using the full dataset would cause for a higher Kaggle score.

- Models Used
  - Random Forest - Bagging/boost strapping aggregating
  - Bayesian Additive Trees - Essentially looped boosting algorithm
  - Logistic Regression
  - Extra Trees - Bagging
- Kaggle AUC: 0.70311
- Runtime: 1.05 hours