KIVA CROWDFUNDING ANALYSIS

APPLYING MACHINE LEARNING TO UNDERSTAND KIVA MICROLENDING AIMING TO ALLEVIATE POVERTY

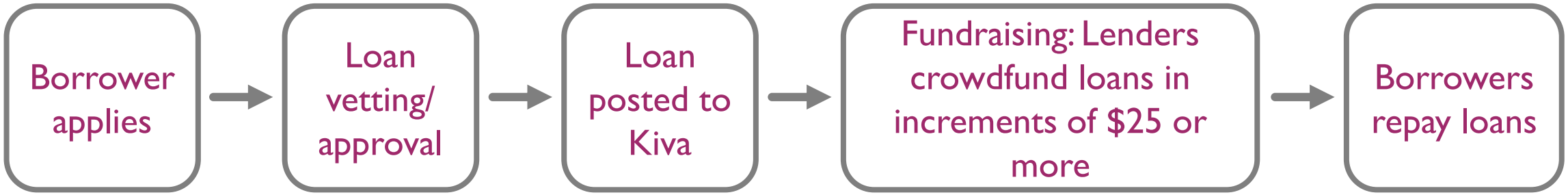
CAROLE HOUSE

Images from www.kiva.org

BACKGROUND

ABOUT KIVA

- **What is Kiva?** An international non-profit hosting a crowdfunding platform that connects prospective borrowers with lenders to receive interest-free capital (up to $10,000)
- **How do Kiva loans typically work?**



ABOUT KIVA

- **Whom does Kiva help?** Millions around the globe.

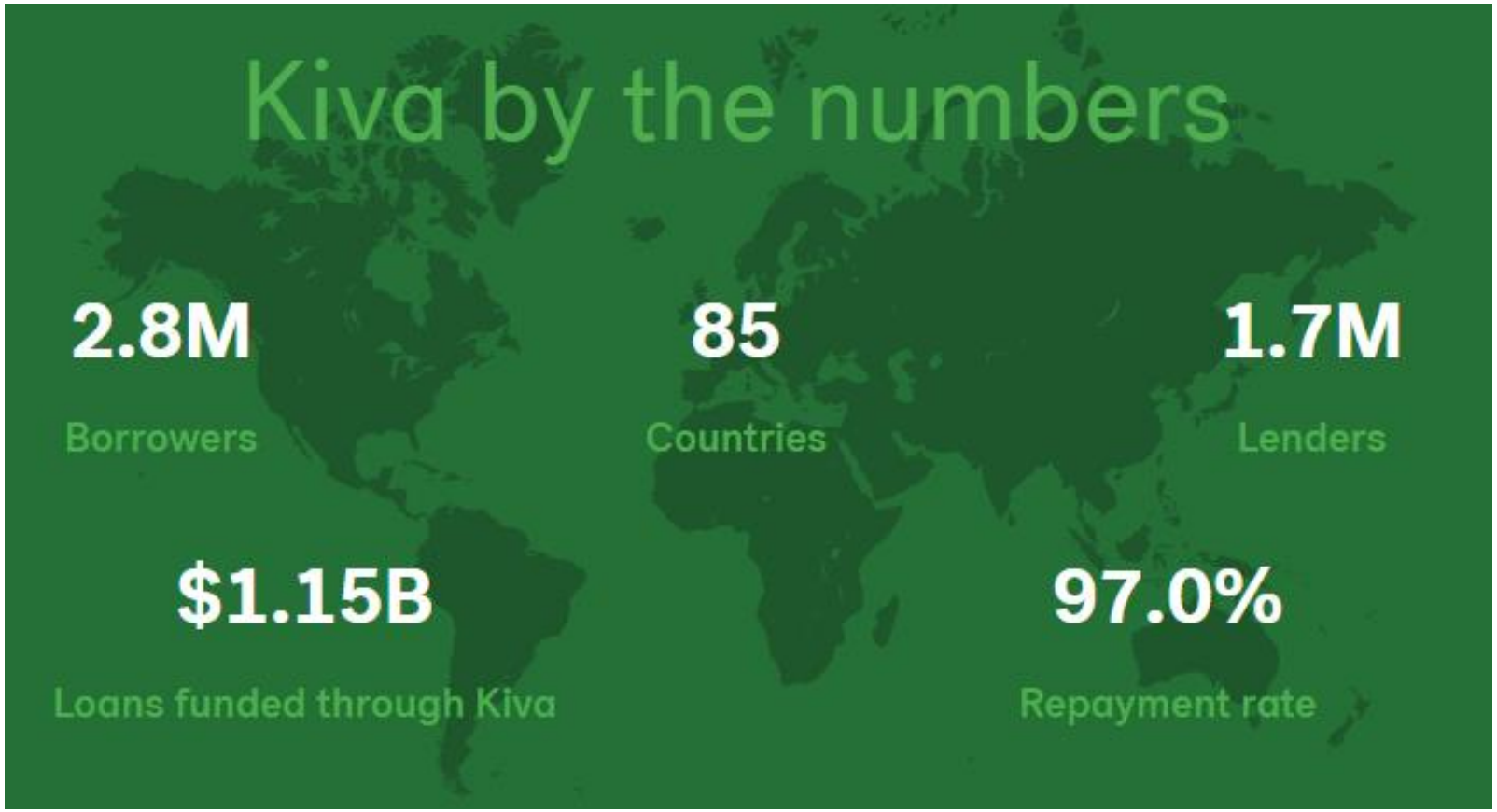


Image from www.kiva.org

KIVA.PRETTY_COOL

(GET IT? BEING PRETTY COOL IS AN ATTRIBUTE OF KIVA! COME ON, THAT'S FUNNY)

Many of us have traveled to lot of countries and seen the effects of poverty and conflict there. The idea of being able to directly connect with and support these entrepreneurs and communities is pretty darn cool.

THE PROBLEM

DATA SCIENCE PROBLEM

Can I accurately predict the loan amount a borrower will receive based on characteristics of the loan and the borrower?

- **Inspiration = Kaggle "Data Science for Good" Competition**
  - **Purpose:** develop machine learning models that can enable more targeted lending and loan application in consideration of borrower poverty levels
  - **Data Provided:** Kiva loan data snapshot (does not include information indicative of borrower poverty level)

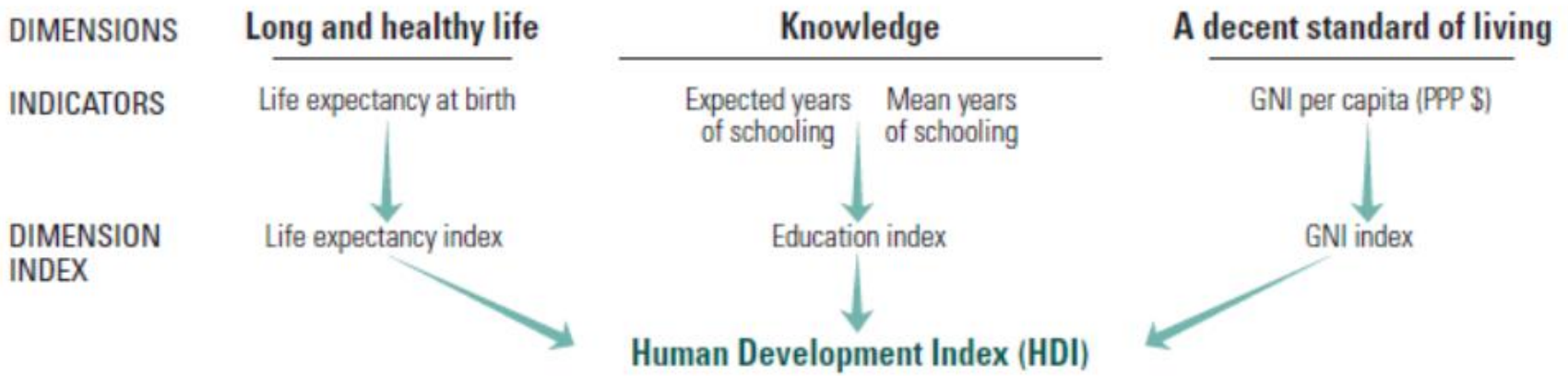
WHY SOLVE THIS PROBLEM?

- **Help borrowers** predict the likely amount they will receive in kiva loans
  - Can assist in managing or honing expectations in business planning
  - Could drive improvements in borrowers' applications if they can affect influential features
- **Provide lenders and kiva representatives visibility** into current lending activity
  - Could drive policy changes if elements like poverty levels don't play as significant a role in affecting lending activity as desired
  - Could change the type or amount of data collected from or about the borrower based on influence in lending activity

THE DATA

DATASETS

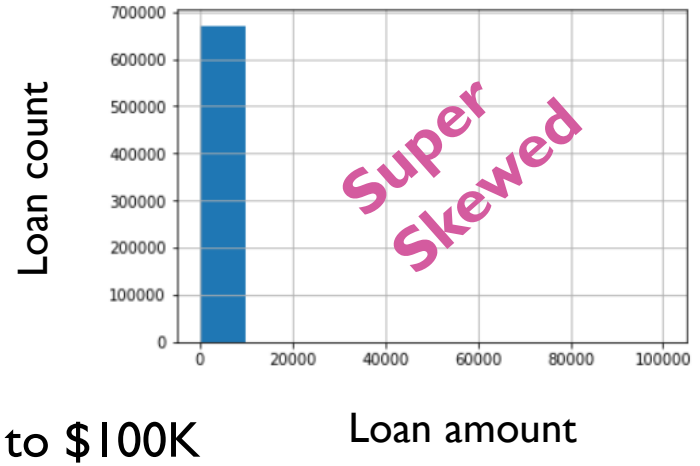
- **Kiva loan data** (since 2014)
- **Armed Conflict and Law Enforcement Database (ACLED) Project data** (since 2014) – a project analyzing and mapping forms, actors, dates, and locations of political violence and protests
- **United Nations Development Programme Human Development Index (HDI) data** (by country, 2014)



Kaggle-Kiva data from https://www.kaggle.com/kiva/data-science-for-good-kiva-crowdfunding/data
ACLED data from https://www.acleddata.com/data/
UNDP data and image from http://hdr.undp.org/en/content/human-development-index-hdi

CLEANING AND INTEGRATING THE DATA

- **Managed nulls**
- **Integrated conflict and development data for each loan**
  - Import ACLED conflict status for loan country – "yes" or "no" for nation inclusion in the database since 2014)
  - Identify and resolve different spellings
- **Managed significant data skew**
  - Limited the project to loans under $1000 (semi-normal)
    - Majority (75%) of kiva microloans are within that range
    - The model for lending would likely be different for loans of up to $100K

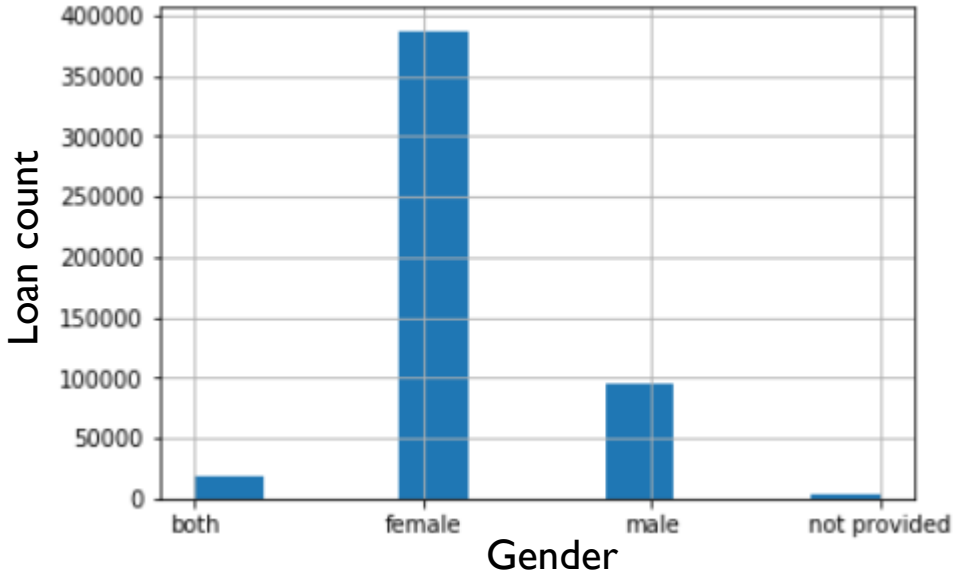


- **Dropped unnecessary columns**

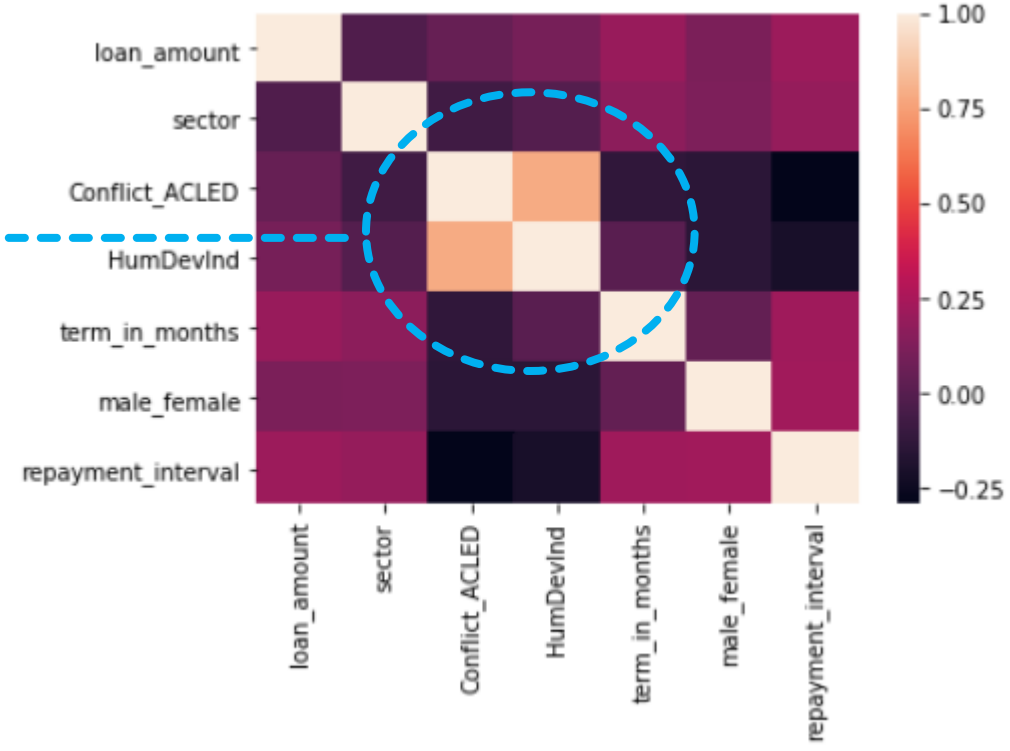
*****Note to self: Don't save your excel doc as a csv first and then lose all your beautiful code and tabs**

EXPLORING THE DATA

- **Top three activity sectors funded:** Agriculture (27%), Food (21%), Retail (20%)
- **Top three countries home to loan recipients:** Philippines (31%), Kenya (14%), El Salvador (6%)
- **Females receive most kiva loans** (77%)
- **Limited correlation** between any single feature and loan amount



Potential interdependency? Remove one of these features?



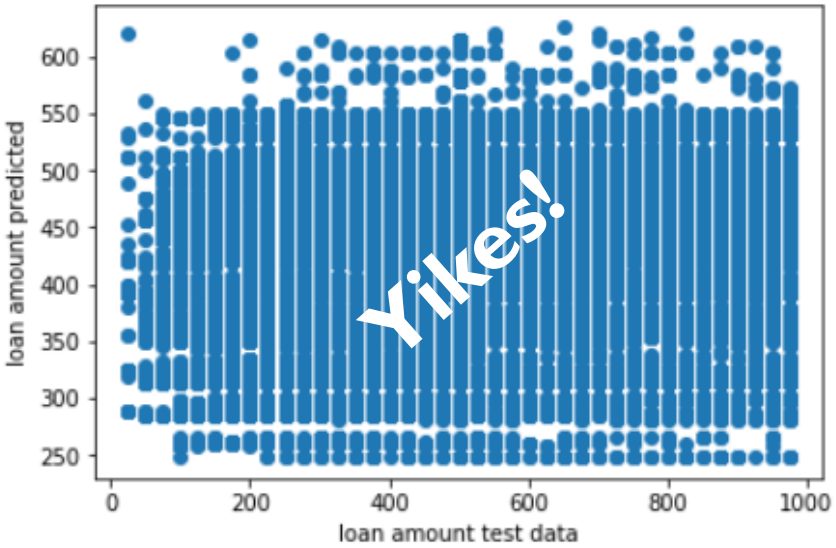
GOAL

- **Predict the amount of a kiva loan** granted to a borrower based on features of the loan and borrower
  - Identify the most important features/predictors of loan amount
  - Target Variables
    - ACLED Conflict Status (categorical, factorized)
    - UNDP Human Development Index
    - Gender of the Borrower(s) (categorical, factorized)
    - Repayment Interval (categorical, factorized)
    - Sector of activity funded by the loan (categorical, factorized)
    - Loan term length

MODELING

LINEAR REGRESSION

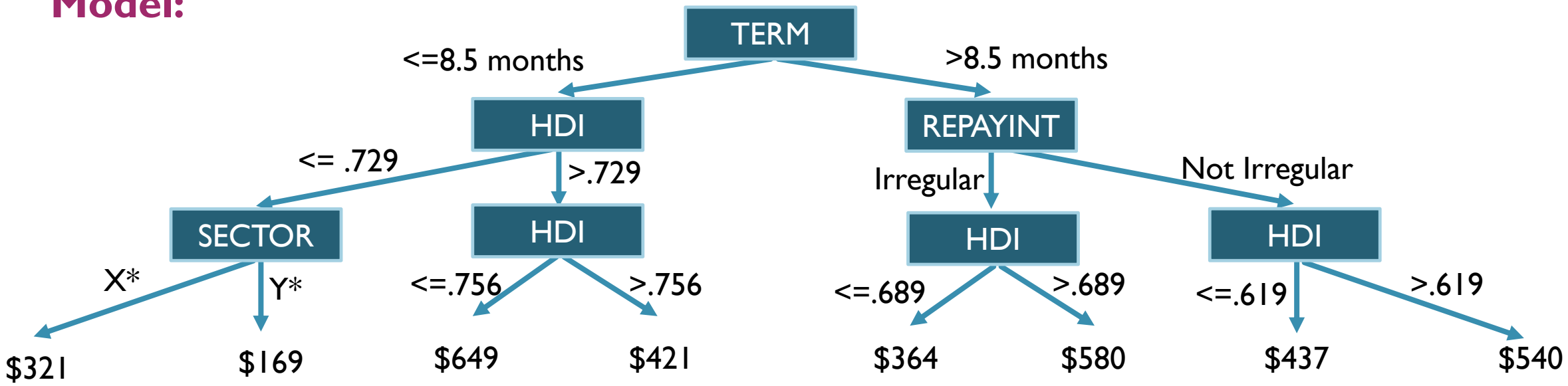
- **Null RMSE (predicting mean):** 225.73
- **Most Important Features** (Recursive Feature Elimination): HDI, Gender, Repayment Interval
- **Model:** loan amount = 429*HDI + 36*Gender + 53*RepaymentInterval + 78
- **Assessment:**
  - Good for interpretability/ease of communication of the impact of variables and maintaining low variance.
  - However, the bias here is very high, resulting in minimal improvement in error from the null RMSE



RMSE: 217.35	BIAS: HIGH	VARIANCE: LOW

DECISION TREE REGRESSION

Model:



X sectors = Food, Transportation, Arts, Services, Agriculture, Manufacturing, Wholesale, Retail, Clothing, Construction, Health

Y sectors = Education, Personal Use, Housing, Entertainment

RMSE: 203.00	BIAS: LOW	VARIANCE: HIGH

DECISION TREE REGRESSION

- **Most Important Features** (Gini importance): Term length, HDI, Repayment Interval
- **Assessment:**
  - In longer-term loans, higher loan amounts went to borrowers in more developed countries
  - Error is less than LinReg
  - My changes for usability (limiting features and tree depth) increased bias and lowered variance
  - Some more improvement in error here, but still pretty high

RMSE: 203.00	BIAS: LOW	VARIANCE: HIGH

RANDOM FOREST REGRESSION

- **Most Important Features** (Gini Importance): HDI, Term Length, Sector
- **Model:** 80 trees in a forest (n_estimators), features at split (max_features)=2
- **Assessment:**
  - Significantly lowered error (though still kind of high), but doesn't provide transparency or interpretability for the impact of certain features
  - RF model has better prediction strength. Shortened DT or LR have been interpretability for communication to policy-makers stakeholders

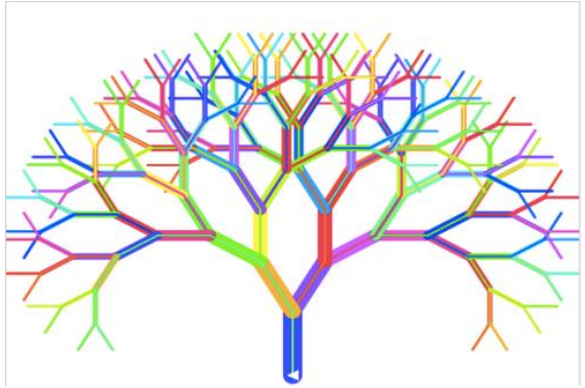


Image from https://medium.com

✓

RMSE: 167.60	BIAS: LOW	VARIANCE: HIGH

INSIGHTS

TAKE-AWAYS

- **None of the models were very satisfying**
  - High RMSE for loans all under $1000 *($167 is a big deal to someone in Somalia where the GDP per capita is ~$430)*
- **HDI** consistently was the **strongest predictor** for loan amount
  - Borrowers should consider the HDI of their nation in applying for loan amounts
  - I can look more closely at other potential relationships between features and HDI
- **Relationships I expected didn't show up** strongly or consistently (e.g., conflict, sector, difference in "important" variables for each model)
- **Opportunities exist** to capture more data and features to better understand kiva lending

FUTURE WORK

NEXT STEPS TO EXPAND/OPTIMIZE MODEL

- Expand target variables to look at removed features
- Integrate more detailed poverty and development indicator data (e.g., the inputs for the HDI scores)
- Refine location information to improve categorization of conflict presence and look at other features (e.g., region, latitude/longitude, proximity to metropolises, etc.)
- Apply classification modeling to answer other questions for this data set (e.g., predicting whether the gender of a loan recipient is male or female when gender is not provided)
- NLP (activity information) and time-series analysis to look for other trends

QUESTIONS?

ONLY EASY QUESTIONS, THOUGH. HARD QUESTIONS NOT ALLOWED.