Capital One

By Joshua Whitaker, De'Johnna Wright, and Caelia Thomas

Advisor: Dr Imtiaz Ahmed

4/15/2022

Problem definition - Background

- Machine learning-driven analysis of complex data sets
- Scenario: Determine rather an account holder is eligible for a loan
- Primary criteria for loans:
  - Annual Income
  - Employment length
  - Verification status

Sample Datasheet

funded_amn	term	int_rate	installment	grade	sub_grade	job_title	emp_length	home_owne	annual_income	verification_status	loan approve	loan purpose
3600	36 months	7.49%	111.97	A	A4	Code/Compl	10+ years	MORTGAGE	120000	Not Verified	No	Other
15000	60 months	14.99%	356.78	C	C4	Senior Super	10+ years	MORTGAGE	125000	Not Verified	Yes	Other
8400	36 months	11.39%	276.56	B	B3	IT Tech	8 years	MORTGAGE	50000	Source Verified	Yes	Other
4000	36 months	10.49%	130	B	B2	Dental hygie	2 years	RENT	50000	Source Verified	Yes	Major purcha
6000	36 months	7.24%	185.93	A	A3	Program Ma	3 years	RENT	125000	Source Verified	Yes	Credit card re
20000	36 months	15.99%	703.05	C	C5	Business de	1 year	RENT	77000	Not Verified	No	Debt consolic
5000	36 months	14.99%	173.31	C	C4	Lead Supervi	10+ years	RENT	68000	Not Verified	No	Debt consolic
20000	36 months	8.24%	628.95	B	B1	IT Lead Busi	10+ years	MORTGAGE	100000	Source Verified	Yes	Debt consolic
12000	36 months	7.99%	375.99	A	A5	registerd nu	9 years	MORTGAGE	100000	Verified	Yes	Debt consolic
9900	36 months	5.32%	298.14	A	A1	Radtech	10+ years	MORTGAGE	53000	Not Verified	No	Debt consolic
10000	36 months	7.49%	311.02	A	A4	Product Ma	4 years	RENT	85000	Not Verified	No	Debt consolic
12000	36 months	6.99%	370.48	A	A2	signmaker/	10+ years	MORTGAGE	56000	Not Verified	No	Home improv
9000	36 months	11.39%	296.32	B	B3	SUPERINTE	10+ years	MORTGAGE	60000	Not Verified	No	Medical expe
5000	36 months	25.49%	200.1	E	E4	Pilot	10+ years	MORTGAGE	215000	Source Verified	Yes	Debt consolic
16000	60 months	12.74%	361.93	C	C1		10+ years	MORTGAGE	72000	Not Verified	No	Major purcha
14400	36 months	10.49%	467.97	B	B2	Property M	6 years	MORTGAGE	130000	Not Verified	Yes	Vacation
6000	36 months	11.44%	197.69	B	B4	Maintenanc	5 years	RENT	25000	Not Verified	No	Debt consolic
20000	36 months	10.49%	649.96	B	B2	driver	10+ years	RENT	95000	Not Verified	No	Other
17000	36 months	7.49%	528.73	A	A4	Air Traffic C	10+ years	MORTGAGE	120000	Not Verified	Yes	Debt consolic
35000	60 months	25.49%	1037.38	E	E4	Human Res	10+ years	MORTGAGE	114000	Source Verified	Yes	Home improv
6000	36 months	8.24%	188.69	B	B1	Lead design	1 year	MORTGAGE	110000	Not Verified	Yes	Credit card re
10150	36 months	7.24%	314.52	A	A3	Compliance	8 years	MORTGAGE	50000	Not Verified	No	Debt consolic
4700	36 months	15.99%	165.22	C	C5	RN Manage	10+ years	MORTGAGE	100046	Source Verified	Yes	Debt consolic
1500	36 months	5.32%	45.18	A	A1	Auto Techn	3 years	MORTGAGE	67000	Not Verified	No	Car financing

Problem Definition

- Initial Goal:

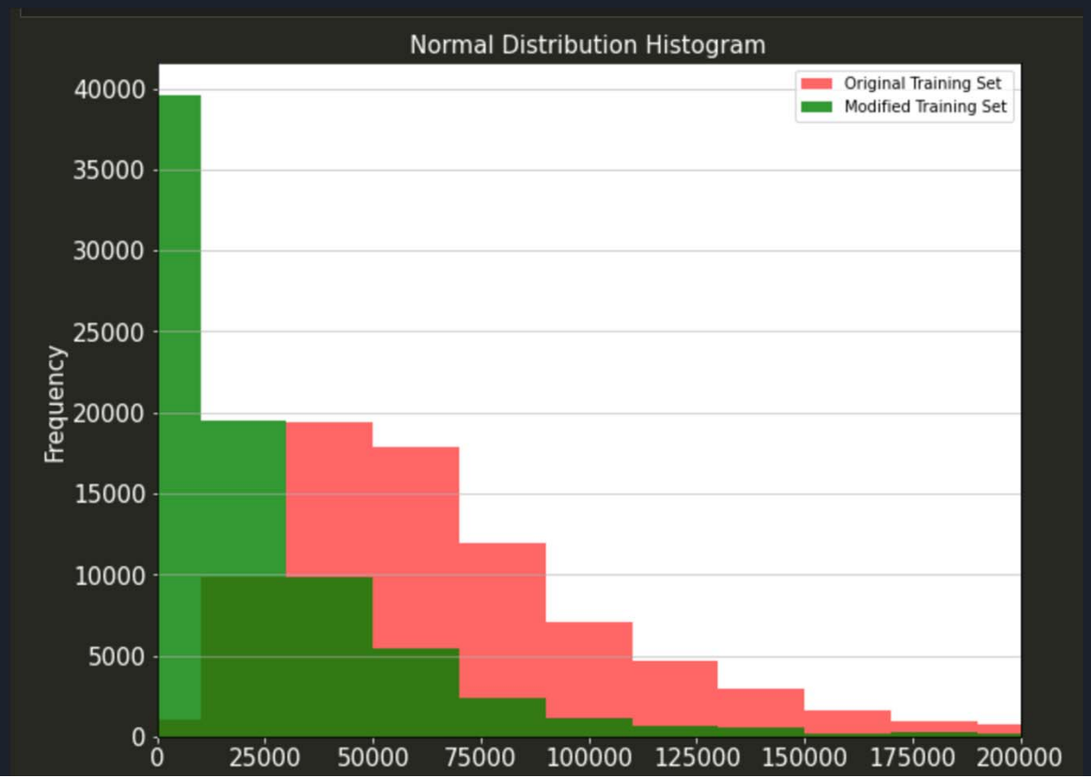
  Time-Varying Regression Model

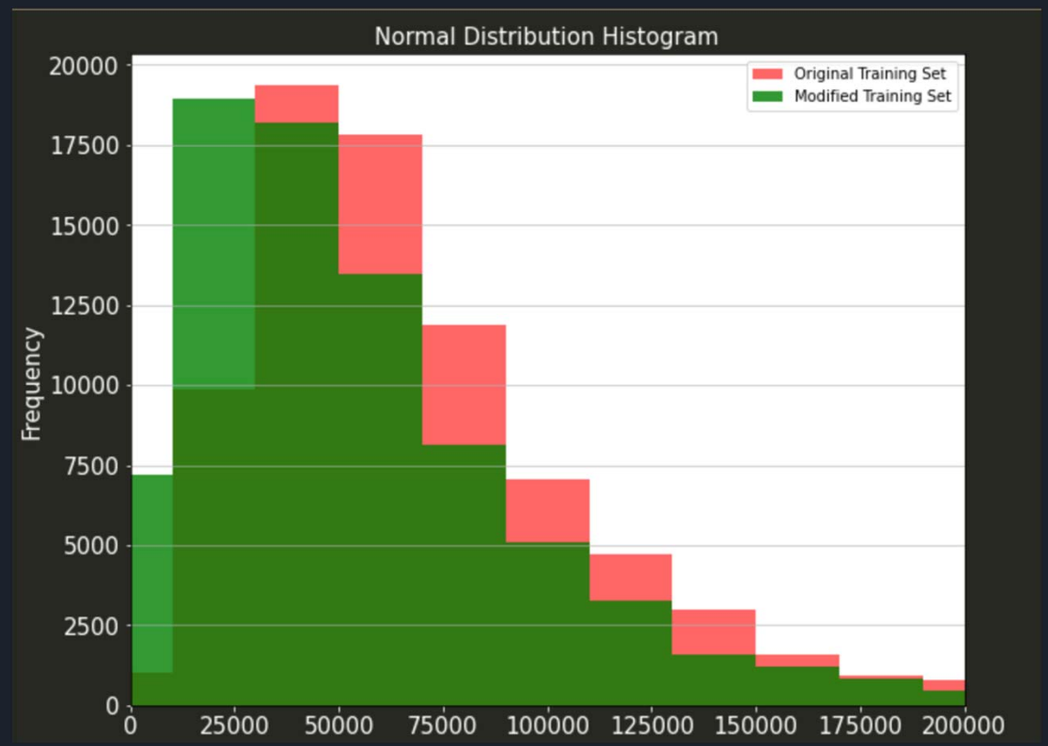
- Problem Statement:

  The need for Capital One is to increase efficiency in processing data sets to determine loan eligibility, and effectively visualize data for in depth analysis by using machine learning algorithms.

What is Data Drift?

- Data drift occurs when values from an input dataset shift away from the values that the dataset that regression model was trained





Design Requirements

Refined Problem Definition:

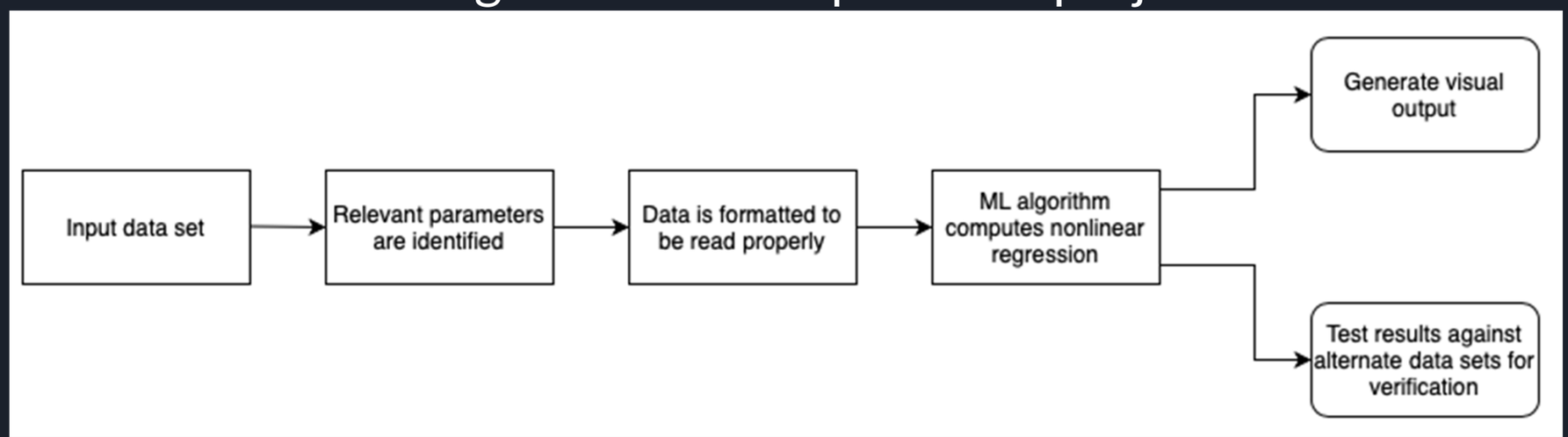
- The need for the Capital One team is to increase efficiency in processing data sets to determine loan eligibility and ensure that the algorithm in place will be accurate even as data changes over time.

Design Requirements

- Constraints
  - Must use machine learning approach
  - Must use strictly open source libraries
  - Time, received project information mid January
- Regulations
  - Report updates to Capital One on monthly basis
  - Comply with Capital One’s NDA

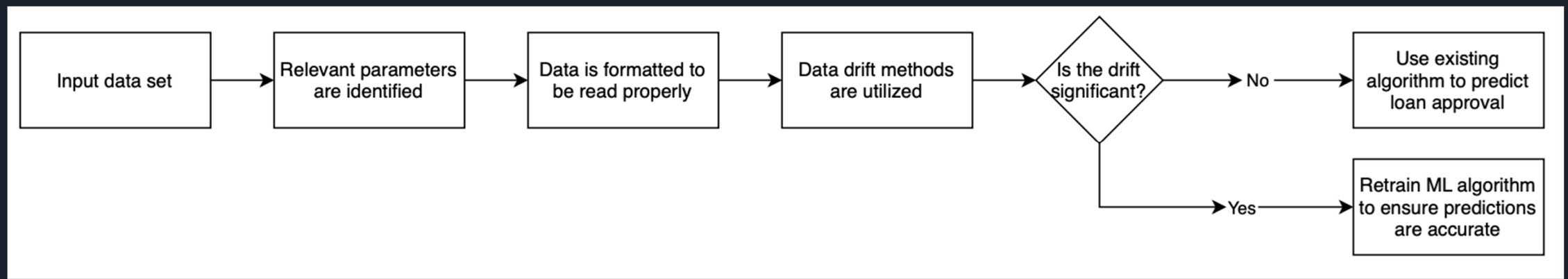
Solution Design

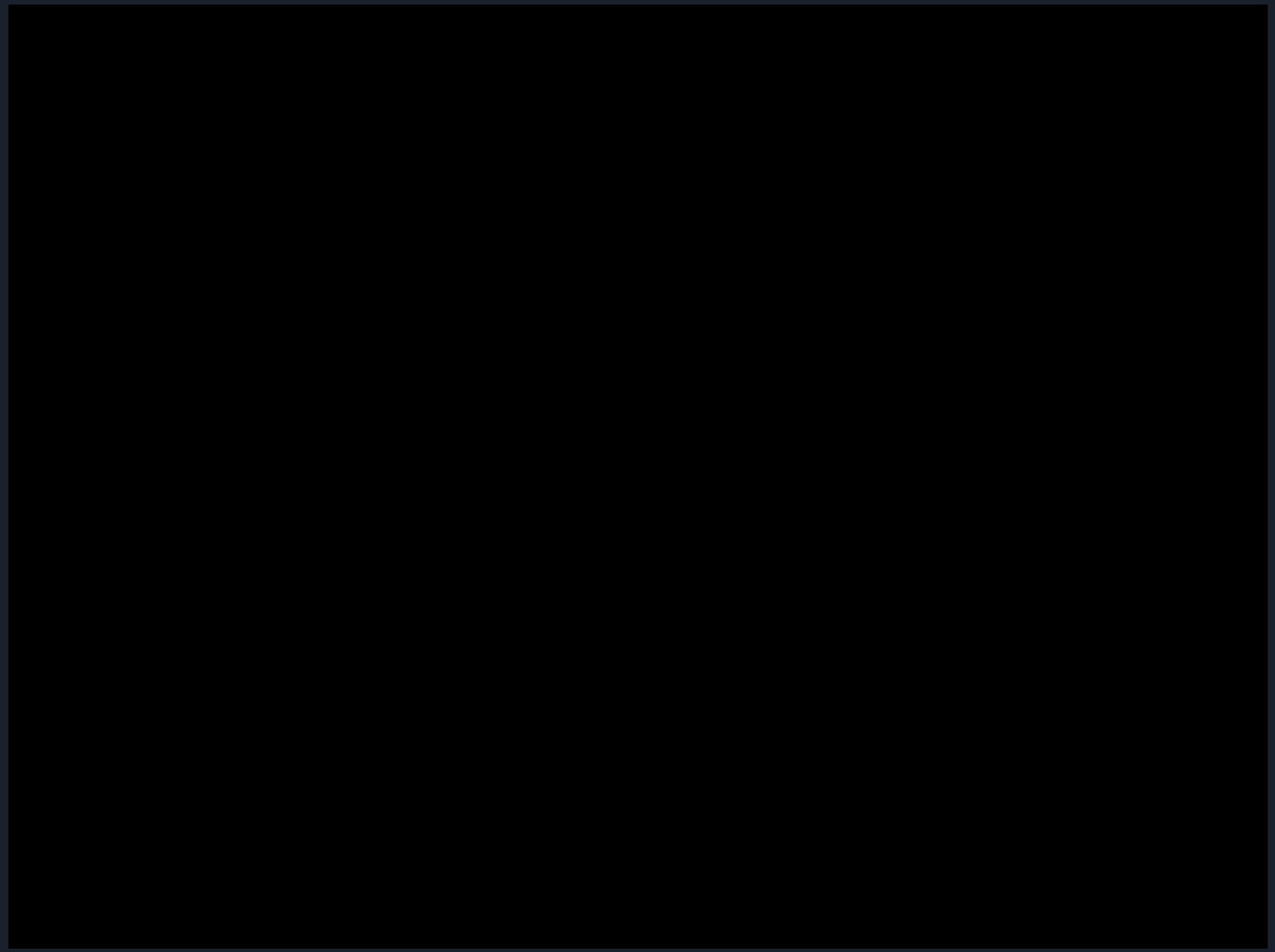
- Initial Solution = fully functioning regression model
- However, project did not officially start until January 2022
  - Not enough time to complete full project



Solution Design

- Solution for Spring 2022
- Focus on data drift detection and integration
  - Next year the project can build off our progress





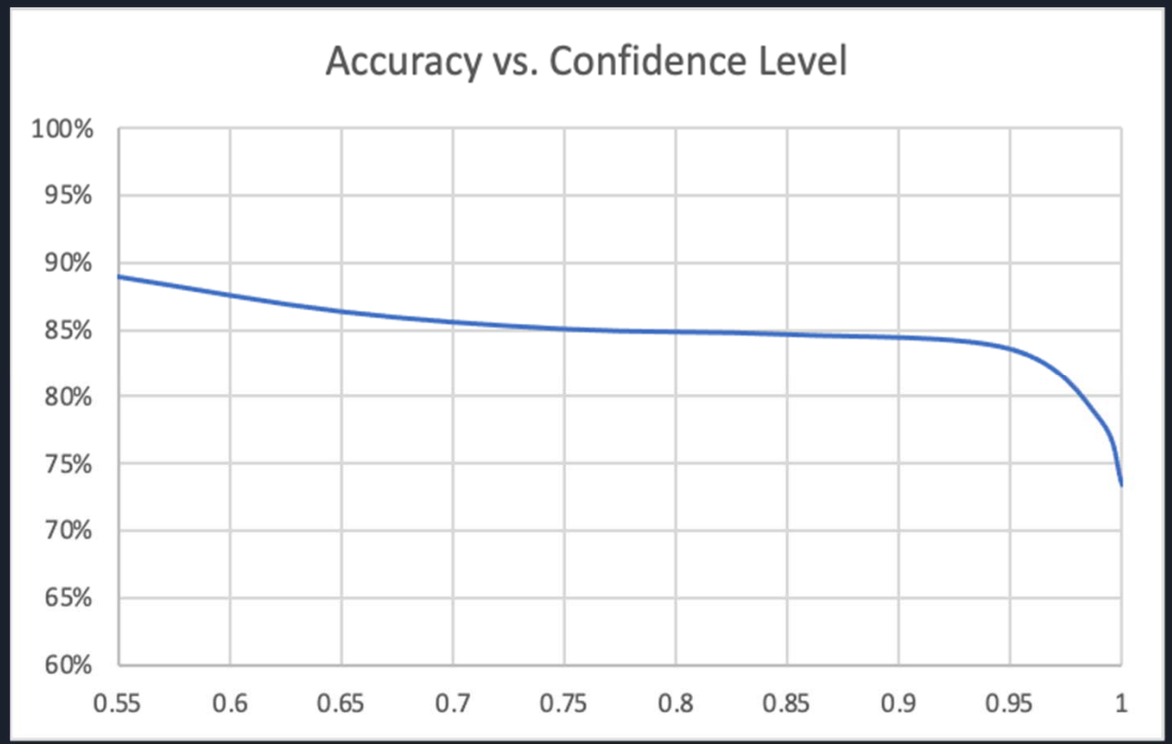
Implementation Process - Sprint #1

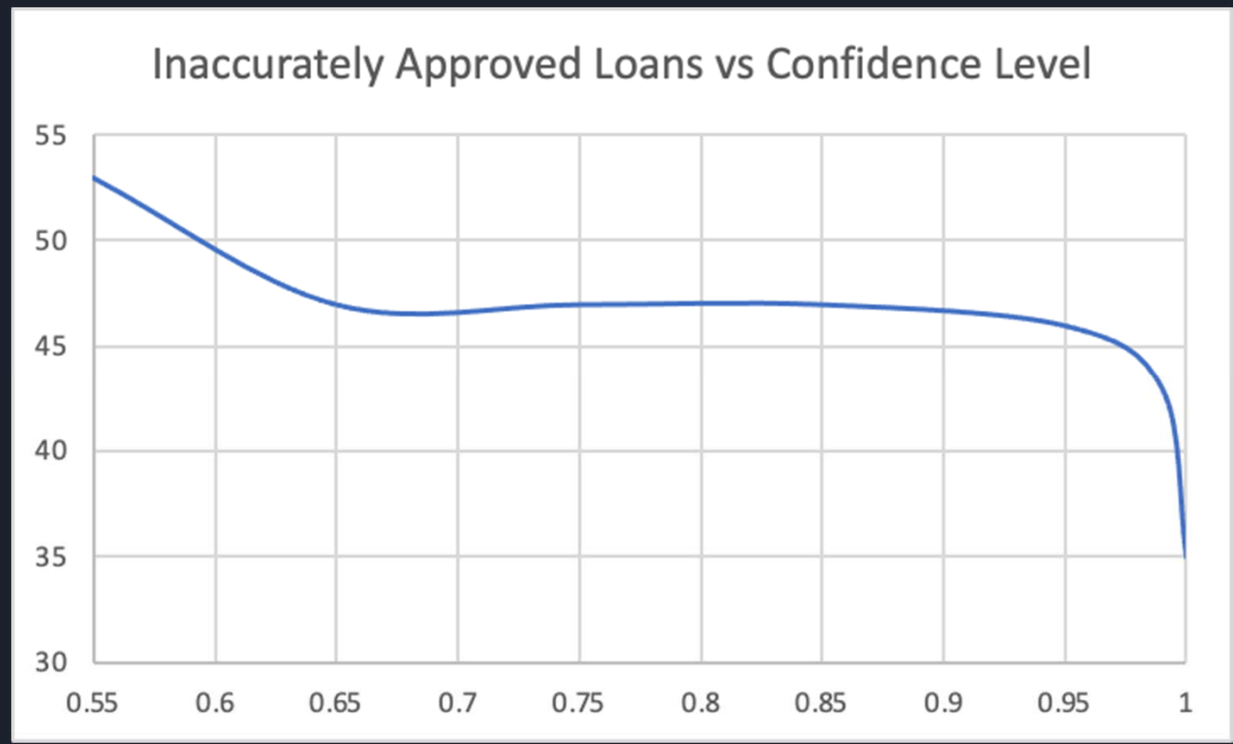
- Determine loan eligibility for account holders
- Cleaned sample datasheet
- Used Python library to create Random Forest Regression algorithm
- Tested algorithm using a test datasheet against training datasheet

verification_status
Not Verified
Not Verified
Source Verified
Source Verified
Source Verified
Not Verified
Not Verified
Source Verified
Verified

verification_status
1
1
2
2
2
1
1
2
3

Implementation Process - Sprint #1
Model Accuracy Confidence





Based on training data sheet, our values were at most 90% accurate when testing a small sample

Implementation Process - Sprint #2

- Applied different data drifting methods to determine how much modified data has shifted
  - Annual income column’s mean shifted by 14000
- Techniques used:
  - PSI
  - Adwin
  - K-S
  - Page Hinkley
  - HDDM_W

Implementation Process - Sprint #2

Method	Results
PSI	Stability index: 0.22
K-S	K-S value: 0

Method	Results
Adwin	7 changes detected
Page Hinkley	2714 changes detected
HDDM_W	18420 changes detected

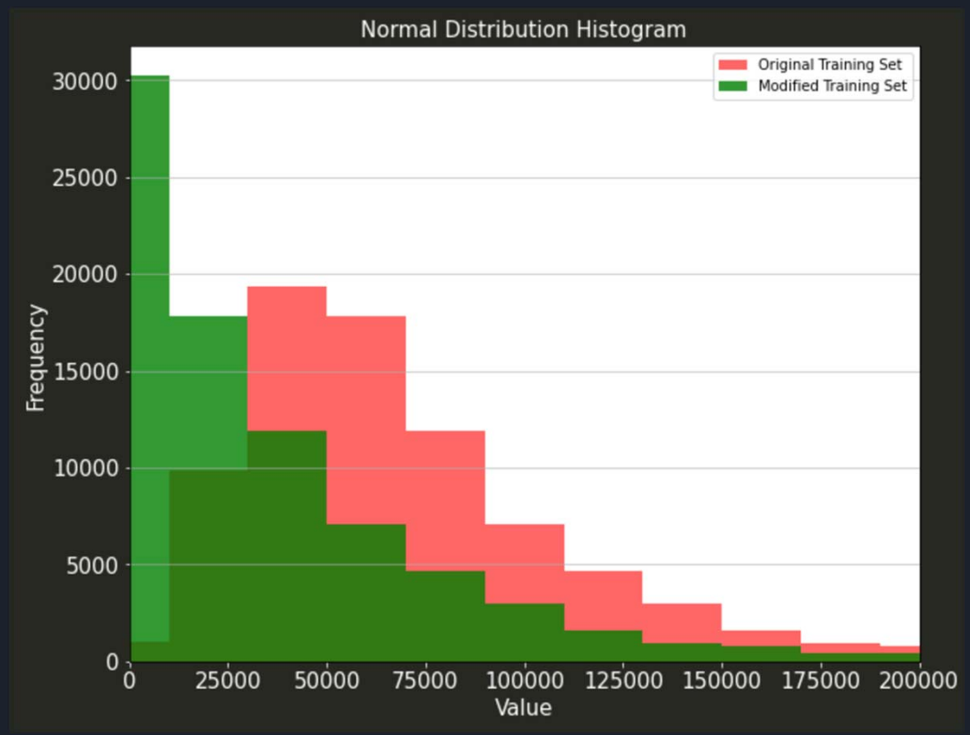
*The results found used only the annual income column of the original and modified dataset.

Implementation Process - Sprint #3

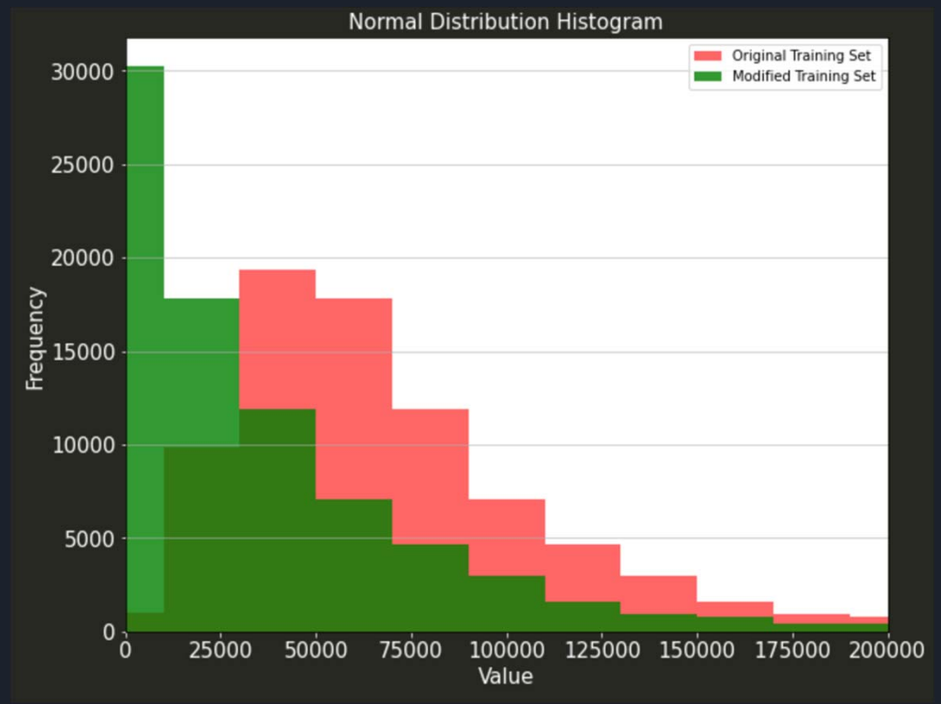
- Based on Phase 2 findings, determined that PSI was the most effective data drift method with the most meaningful results
  - The other methods vary too much in results to be meaningful
- Continued to test the methods to confirm if they worked by using four additional modified datasheets

Implementation Process - Sprint #3

Dataset 2
Mean of annual income reduced by 40000
PSI: 1.1

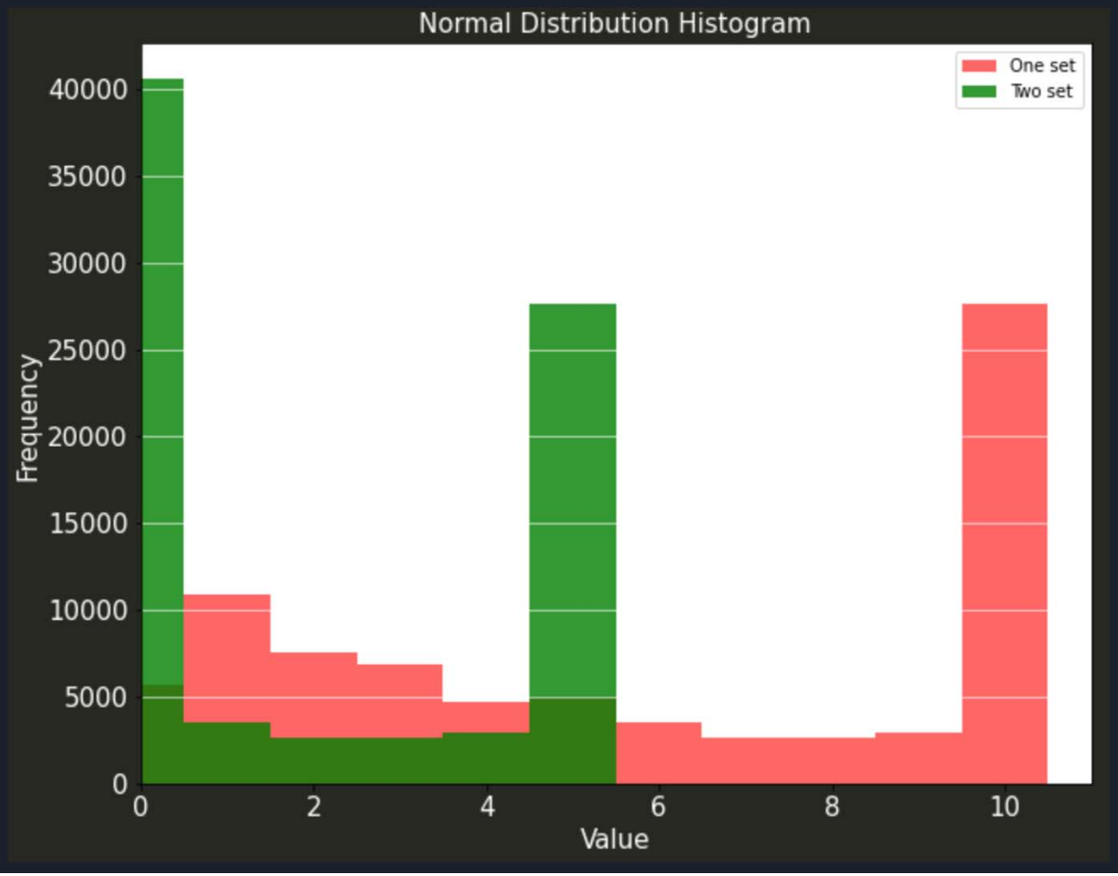


Dataset 3
Mean of annual income reduced by 40000
Variance reduced by half
PSI: 1.77



Implementation Process - Sprint #3

Dataset 4
Employment length decreased by 5 years
PSI: 5.25



Implementation Process - Sprint #4

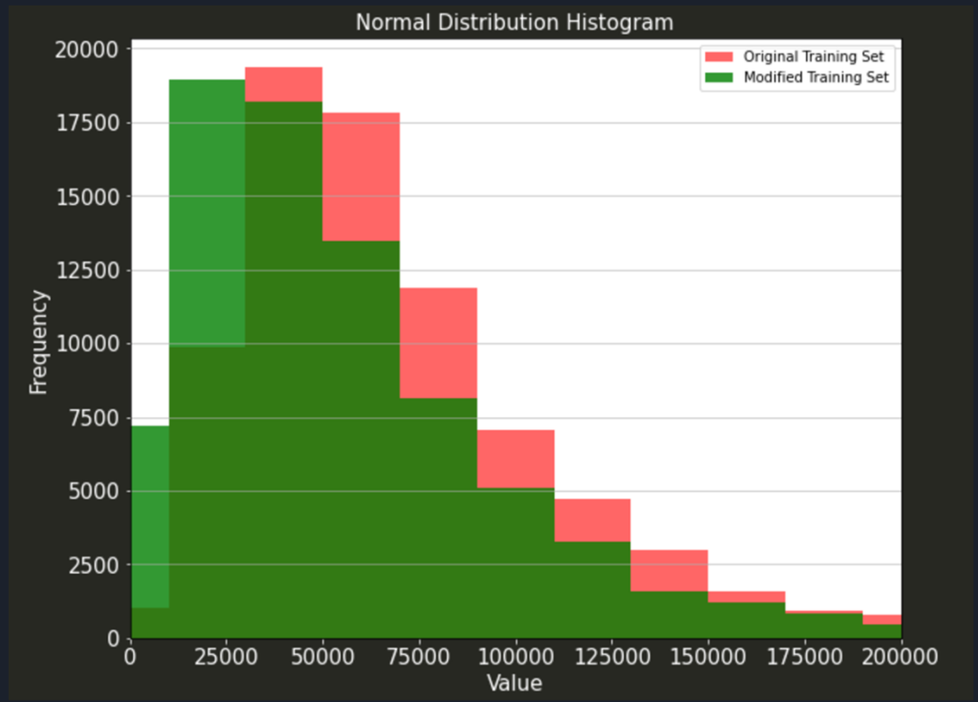
- In addition to data drift, started taking concept drift into account
  - Concept drift is how much the algorithm's predictions are affected by drifted datasets
- Ran original algorithm against modified datasets to evaluate how the drifted data affected the random forest predictions

Implementation Process - Sprint #4
Concept Drift Confidence

Dataset	PSI	Accuracy (%)	Inaccurately Approved Loans
1	0.22	82.1	45
2	1.1	69.89	31
3	1.81	66.7	28

*All modified datasets' results were compared to the original dataset' results

Conclusion




```
regressor = RandomForestRegressor(n_estimators = 100, random_state = 0)

# fit the regressor with independent and dependent variables
regressor.fit(df, df_approved)
```

Member ID	Loan Approved	Correctly Predicted
80001	1	TRUE
80002	1	TRUE
80003	1	TRUE
80004	1	TRUE
80005	0	FALSE
80006	0	FALSE
80007	1	TRUE
80008	1	TRUE
80009	0	TRUE
80010	0	FALSE
80011	1	TRUE
80012	1	TRUE
80013	1	TRUE
80014	1	TRUE

Thank You!