



# Individuals Annual Income in The USA

---

Yasir Albahlal, Turki Almuarik

# Content

- Project Scope
  - Data
  - EDA
  - Baseline
  - Models
  - Balancing and Feature Engineering
  - Graphs
  - Conclusion
-

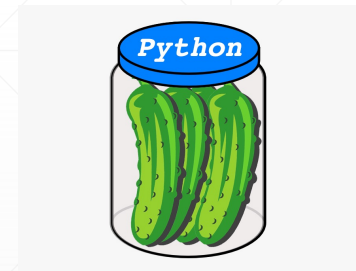
## Project Scope:

- The goal of our project is to build a classification model to predict whether a US resident annual income is more or less than 50k.

## Tools:

matplotlib

pandas



# Data

- Data imported from the US census website
- Contain almost 300k rows
- more than 40 columns
- Columns such as age, wage per hour, gender, weeks worked, etc.



# EDA

- Removing duplicates
  - Removing null values
  - Removing columns
  - Striping values
  - Changing target to 0 and 1
  - Rows after EDA: 285k
  - Columns after EDA: 27
-

# BASELINE

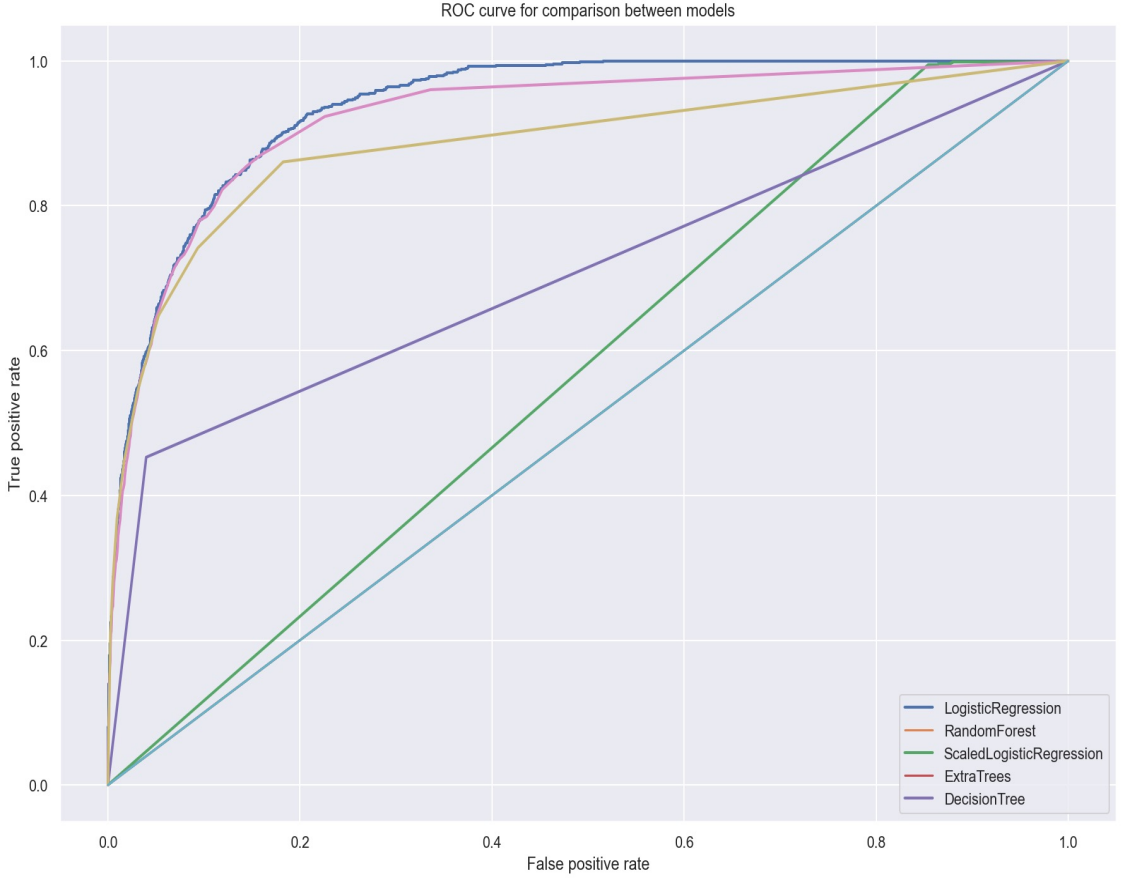
- kNN is the baseline model
- Building a model at the very beginning
- Model overfitting

Model name	Training	Validation	F1 scores	Precision	Recall
kNN baseline	0.961	0.937	0.391	0.481	0.330

# Models

Model name	Training	Validation	F1 scores	Precision	Recall
kNN baseline	0.961	0.937	0.391	0.481	0.330
kNN	0.948	0.939	0.294	0.734	0.184
Logistic Regression	0.950	0.946	0.460	0.747	0.332
Scaled Logistic Regression	0.950	0.948	0.498	0.755	0.372
Decision Tree	0.999	0.925	0.453	0.455	0.452
Extra Trees	0.999	0.945	0.493	0.668	0.391
Random forest	0.994	0.947	0.491	0.743	0.367

# ROC



# Balancing and Feature Engineering

## Logistic regression

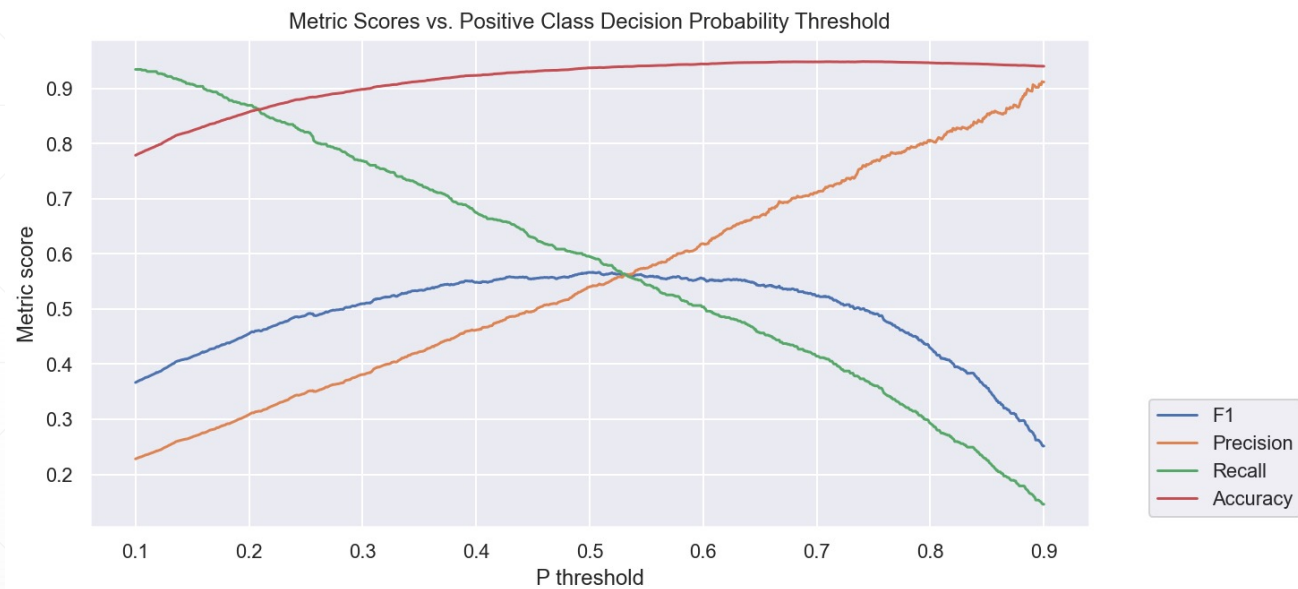
- Balancing data by 3 \* 1
- Trying different feature engineering
- Squaring the age
- Using stacking

Model name	Training	Validation	F1 scores	Precision	Recall
Logistic Regression	0.950	0.946	0.460	0.747	0.332
Logistic Regression 3*1	0.939	0.937	0.564	0.538	0.593
Logistic Regression 3*1 age^2	0.938	0.937	0.567	0.541	0.595
stacked	0.951	0.949	0.510	0.768	0.382

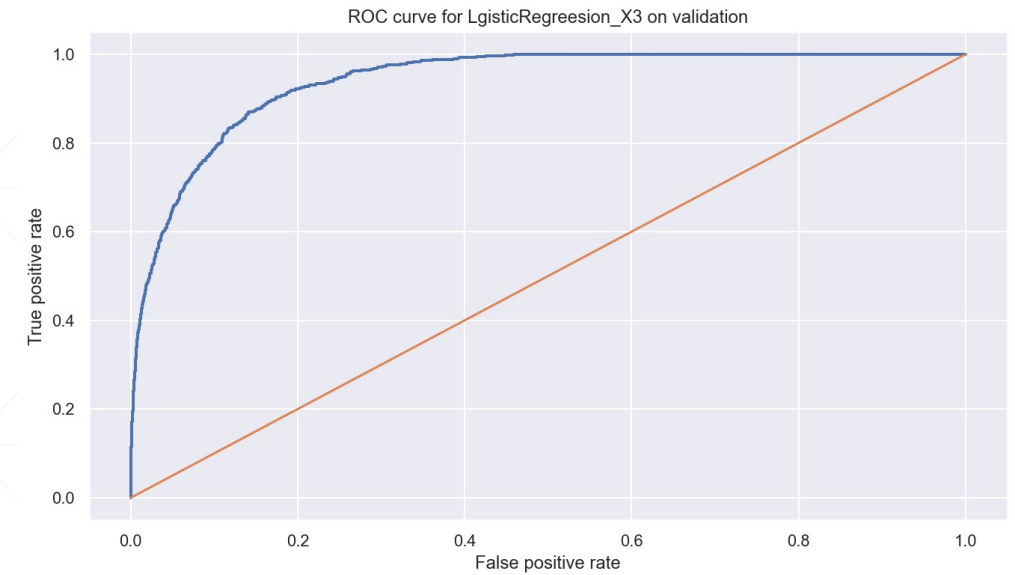


# Graphs: Logistic Regression

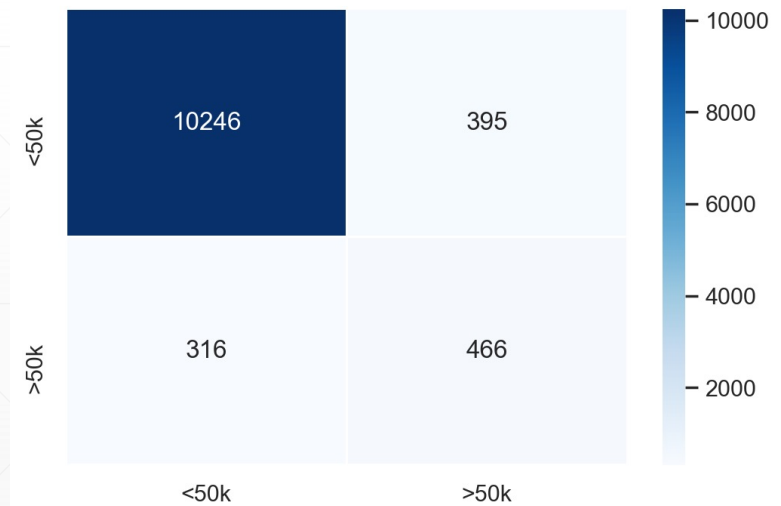
Best threshold at 0.526



## ROC



## Confusion matrix



# Conclusion

## Testing selected model on testing data

- Combining train and validation data
- Training the model
- Using testing data to get scores

Model name	Training	Testing	F1 scores	Precision	Recall
Logistic Regression	0.938	0.939	0.544	0.520	0.569

# Thank you

*Any question?*

---