BLUE BOTTLE COFFEE

Pt 2: Analysis

BLUE BOTTLE
COFFEE

Alondra Espinoza

Agenda

- Introduction
- Problem Statement
- Data Description and Preparation
- Model Building and Evaluation
- Key Findings and proposals
- Concerns and Resolutions

Hello!

Alondra Espinoza

Business Data Analytics

International Business Studies

PROBLEM STATEMENT

Problem:

- Low market share
  - ~99 locations
  - Starbucks: 38,038
  - Dunkin Donuts: 13,200

Proposal:

- Analyze Salary prediction data to identify high paying costumers and where they reside
- Decision tree, Logistic regression

Why:

- Posesses a Luxury atmospshere

Data Description and Preparation

Kaggle: Salary prediction data

Google Colab

7 columns, 1000 records:

- Education, Experience, Location, Job title, age, gender, salary

```
import pandas as pd

# Adjust this path with the correct filename from the directory
file_path = "/content/dataset/salary_prediction_data.csv"
df = pd.read_csv(file_path)
print(df.head())
```

```
     Education  Experience  Location Job_Title  Age  Gender         Sa
0  High School           8     Urban   Manager   63    Male   84620.05
1          PhD          11  Suburban  Director   59    Male  142591.25
2     Bachelor          28  Suburban   Manager   61  Female   97800.25
3  High School          29     Rural  Director   45    Male   96834.67
4          PhD          25     Urban   Analyst   26  Female  132157.78
```

```
print(df.shape)
```

```
(1000, 7)
```

Preparation

- Removed: Experience
- Change salary to integer
- New column “Salary_Category”
- Null values

```
import pandas as pd
import numpy as np

# Create a new column for Salary with standardized labels
df['Salary_Category'] = np.where(df['Salary'] >= 100000, '>=100k', '<100k')
print(df.head())

# Check the result to ensure uniform labels
# print(df[['Salary', 'Salary_Category']].head())
```

```
     Education  Location Job_Title  Age  Gender  Salary Salary_Category
0  High School     Urban   Manager   63    Male   84620          <100k
1          PhD  Suburban  Director   59    Male  142591         >=100k
2     Bachelor  Suburban   Manager   61  Female   97800          <100k
3  High School     Rural  Director   45    Male   96834          <100k
4          PhD     Urban   Analyst   26  Female  132157         >=100k
```

```
] df.head(20)
```

	Education	Location	Job_Title	Age	Gender	Salary	Salary_Category
0	High School	Urban	Manager	63	Male	84620	<100k
1	PhD	Suburban	Director	59	Male	142591	>=100k
2	Bachelor	Suburban	Manager	61	Female	97800	<100k
3	High School	Rural	Director	45	Male	96834	<100k
4	PhD	Urban	Analyst	26	Female	132157	>=100k
5	PhD	Rural	Director	27	Female	156312	>=100k

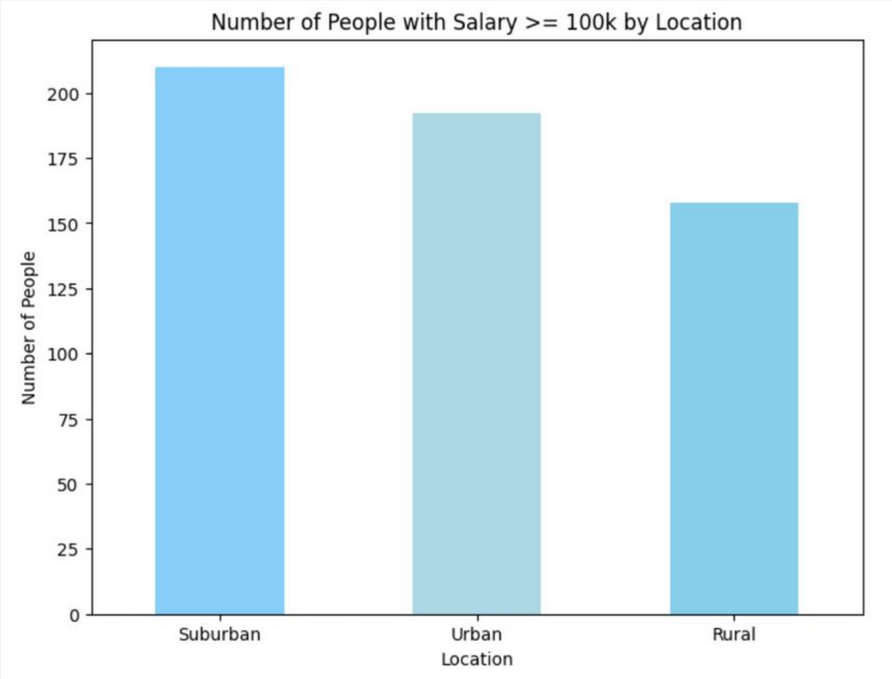
Identify features and target variable

- dummy variables for categorical predictors
- Encode target variable

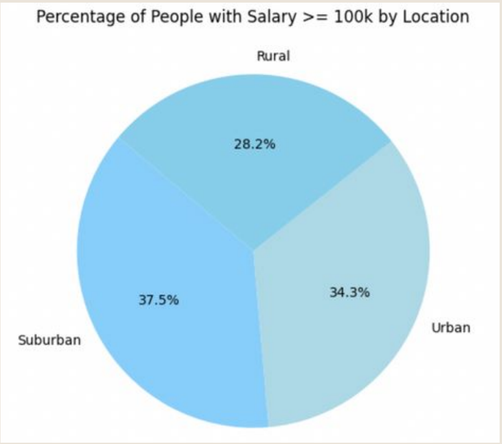
```
#Encoding target variable
from sklearn.preprocessing import LabelEncoder
labelencoder = LabelEncoder()
y = labelencoder.fit_transform(df['Salary_Category'])
y
```

```
array([0, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0,
       1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1,
       1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1,
       1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 1,
       1, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1,
       1, 0, 0, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0,
       1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0,
       0, 0, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0,
       1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 1,
```

Data Exploration



Salary_Category	count
>=100k	560
<100k	440




```
Job_Title
Director    275
Analyst     255
Manager     241
Engineer    229
Name: count, dtype: int64
Location
Suburban    345
Rural       345
Urban       310
Name: count, dtype: int64
```

CLASSIFICATION ANALYSIS

Output variable (Salary_Category) is categorical (<100k, >=100K)

MODELS

- Decision Tree
  - Pros: Easy to interpret, handles both numerical and categorical input variables, Can be improved (Random Forest, Bagging, Boosting)
  - Cons: Prone to overfitting, unbalanced data can lead to bias
- Logistic Regression
  - Pros: Easy to interpret, efficient and accurate
  - Cons: Limited to binary classification, limited to linear relationships

Model Performance

Decision tree

Training set accuracy: 99%

Test set: 85%

Actual/Predicted	<100k	=>100k
<100k	392	1
=>100k	1	500

Model Performance

Decision tree

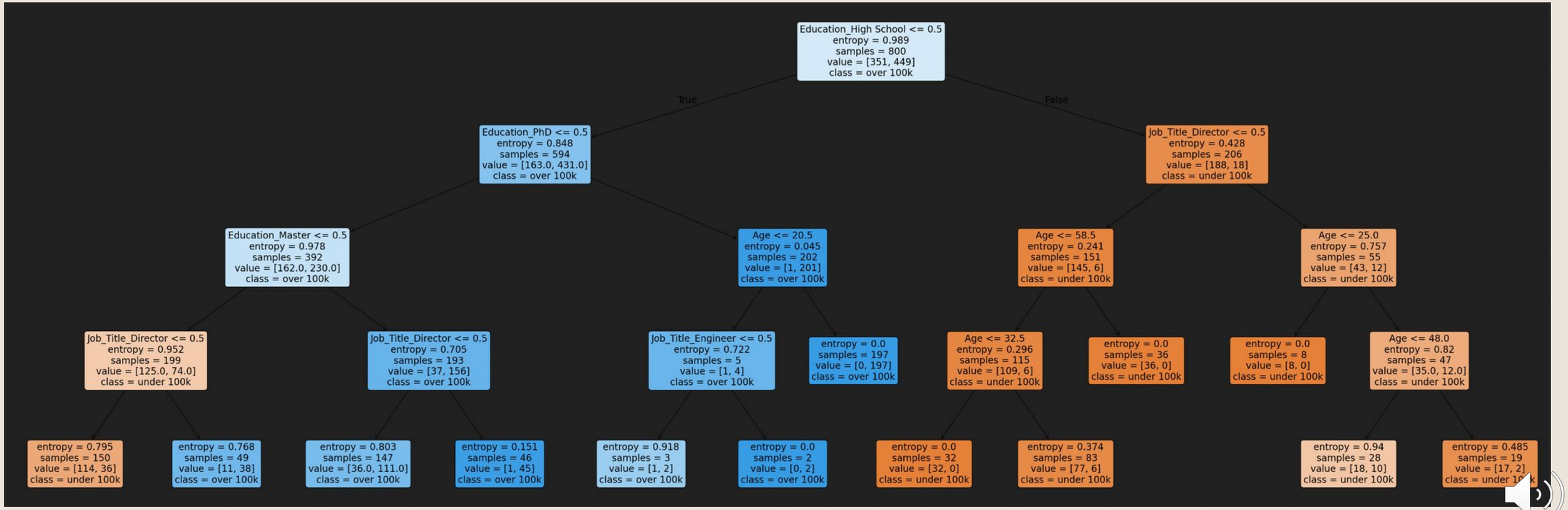
Training set accuracy after pruning: 86%

Actual/Predicted	<100k	=>100k
<100k	39	8
=>100k	7	46

Decision Tree

Education_High School <= 0.5
entropy = 0.989
samples = 800
value = [351, 449]
class = over 100k

- Those who posses a highschool diploma are more likely to achive a salary of or more than 100k
- Director and Phd



Decision Tree

	Importance
Education_High School	0.453773
Education_PhD	0.252261
Education_Master	0.131606
Job_Title_Director	0.122038
Age	0.038380
Job_Title_Engineer	0.001942
Location_Suburban	0.000000
Location_Urban	0.000000
Job_Title_Manager	0.000000
Gender_Male	0.000000

Top Feature Importance

- Education_High School
- Education_PhD
- Education_Master

Model Performance

Logistic Regression

Training set accuracy: %87

Test set: 85%

Actual/Predicted	<100k	=>100k
<100k	260	43
=>100k	48	349

Logistic Regression

	Odds
Education_PhD	13.383412
Job_Title_Director	4.268618
Education_Master	3.147870
Job_Title_Manager	2.191963
Location_Urban	1.989866
Location_Suburban	1.445671
Job_Title_Engineer	1.259980
Age	0.983812
Gender_Male	0.971556
Education_High School	0.354086

Top Feature Importance

- Education_PhD
- Job_Title_Director
- Education_Master

Improved Model Improvement

Not all of these methods increase accuracy, but they should help reduce overfitting

Decision Tree

Test set accuracy: %85

Bagging: 86%

Boosted model: 84%

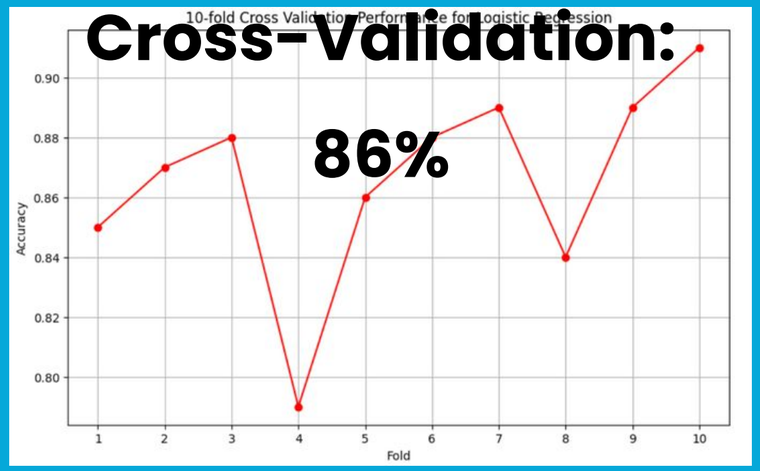
Random Forest: 82%

Logistic regression

Test set accuracy: %85

K-fold

Cross-Validation: 86%



Key Findings and Proposals

- **Education related feature**
  - Location near schools(Universities)
  - Student discount/special

CONCERNS/RESOLUTIONS

- One hot encoding issue
  - Ordinal encoding
- Dataset was limited
  - Access better dataset
- Other analysis
  - Foot traffic analysis

Thank you

for listening!

Sources Used

- Fonts
  - TAN Mon Cheri
  - **Poppins Medium**
  - Poppins Light
  - The Seasons
- Google Colab
- Kaggle
  - https://www.kaggle.com/datasets/mrsimple07/salary-prediction-data
- Website
  - https://bluebottlecoffee.com/us/eng/shop/coffee